ROBUST STATISTICAL METHODS FOR CAUSAL DISCOVERY IN ONE-SAMPLE MENDELIAN RANDOMIZATION STUDIES

By

Ruxin Shi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Doctor of Philosophy

2025

ABSTRACT

Mendelian Randomization (MR) has become a cornerstone approach for inferring causal relationships in epidemiological and genetic studies by leveraging genetic variants as instrumental variables (IV). Despite its popularity, conventional MR analyses, particularly those based on two-stage least squares (TSLS) and conducted within a single sample, face significant methodological challenges. These include selection-induced winner's curse and the pervasive problem of weak instruments and invalid IVs, all of which can undermine the reliability and interpretability of causal effect estimates.

To address these limitations, this dissertation develops a unified and robust MR framework through a sequence of methodological innovations. First, we introduce MR-SPLIT, a novel adaptive sample-splitting and cross-fitting procedure that effectively mitigates biases arising from IV selection and weak instruments in one-sample MR settings. MR-SPLIT employs multiple sample splits to further enhance robustness, demonstrating superior performance in bias reduction, type I error control, and statistical power compared to existing approaches, as validated in extensive simulation studies and real-world data applications. Building on this foundation, we further propose MR-SPLIT+, which integrates best subset selection to accommodate invalid IVs under a relaxed plurality rule. MR-SPLIT+ substantially reduces estimation bias due to invalid instruments while maintaining efficiency and robustness. Simulation results consistently demonstrate that MR-SPLIT+ outperforms contemporary methods, and real-data analyses confirm its practical reliability in complex genetic architectures. Recognizing that causal relationships are often bidirectional or ambiguous, especially within gene expression networks and complex traits, we extend this framework to BiMR-SPLIT+. This method is specifically designed to disentangle bidirectional causality between pairs of traits, even when the underlying IV assumptions are partially violated. Extensive simulation studies and application to Drosophila melanogaster data illustrate that BiMR-SPLIT+ not only recapitulates established biological mechanisms, but also identifies novel candidate genes with potential regulatory roles. This bidirectional MR framework enables more accurate inference of gene-trait relationships and has broad implications for precision medicine.

Collectively, this dissertation presents a cohesive suite of MR methodologies that systematically

address weak and invalid IVs, IV selection bias, and bidirectional causality. The resulting toolkit substantially advances the reliability of causal inference in genetic epidemiology and lays the groundwork for future exploration in complex causal networks as large-scale human datasets continue to grow.

ACKNOWLEDGEMENTS

The five years of my doctoral journey have passed more quickly than I ever imagined. During this time, I traveled to many places, met countless people, and was fortunate enough to solve a few challenging problems. It was a radiant five years, filled with laughter and tears. I am grateful for every moment when, after falling down, I found the strength to begin again.

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Yuehua Cui. Without his guidance and support, this dissertation would not have been possible. Whenever my research reached an impasse, he always helped me discover a new path forward. It has been one of the greatest privileges of my life to be his student.

I also want to thank my boyfriend, Zhouyu Shen, for his unwavering companionship over these five years. In a foreign country, during the height of a global pandemic, he took care of me in countless ways. We made the most of our time together, traveling to many wonderful places during holidays. His presence made these years exceptionally bright.

I am deeply grateful to my parents, Meizhen Lei and Haifeng Shi, for their constant support. As their only daughter, I thank them for backing my decision to pursue a PhD abroad, even though it meant being far from home.

Lastly, I want to thank my two cats, Guoguo and Meimei, whose adorable presence has brought endless comfort and joy to my daily life. I am also deeply grateful for the music of Chenyu Hua, which has greatly enriched my spiritual world.

The past is now behind me—it once accompanied the years of my youth. The present and the future lie ahead, and the world is still vast.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

MR	Mendelian Randomization
IV	Instrument Variable
TSLS	Two Stage Least Square
2SLS	Two Stage Least Square
OLS	Ordinary Least Square
LD	linkage disequilibrium
LIML	Limited information maximum likelihood
IVW	Inverse-variance Weighted
CFI	Cross-Fitted Instrument
SNP	Single Nucleotide Polymorphism
SIS	Sure Independence Screening
CP	Coverage Probability
MAF	Minor Allele Frequency
MIO	Mixed Integer Optimization
FNR	False Negative Rate
FPR	False Positive Rate
DAG	Directed Acyclic Graph

CHAPTER 1

BACKGROUND AND MOTIVATION

1.1 Mendelian Randomization: An Overview

Mendelian Randomization (Davey Smith and Hemani, 2014; Lawlor et al., 2008; Greenland, 2000; Davey Smith and Ebrahim, 2003) is a method using genes as instrument variables (IV) to make causal inference between correlated variables, especially between behavioural, pharmacological or physiological measures and disease. It aims to detect the presence of causal effects and, when present, to provide unbiased estimates of their magnitude. The use of instrumental variables for detecting causal effect is first proposed in Econometric to deal with endogenous variable (Barro, 1997; Wainwright et al., 2005), which is synonymous with a dependent variable and correlates with other factors within the system being studied, and has been discussed a lot during recent years. Since there are some restrictions for the choice of instrumental variables, it is usually not easy to find a perfect one (Donald and Newey, 2001; Baiocchi et al., 2014). However, unlike other variables, people's genotypes are determined only by their parents' genotypes according to Mendel's law (Castle, 1903) and are generally unrelated to those confounding factors that distort the interpretations of findings from observational epidemiology. Furthermore, disease processes do not alter germline genotype and therefore associations between genotype and disease outcomes cannot be affected by reverse causality. Finally, genetic variants that are related to a modifiable exposure will generally be related to it throughout life from birth to adulthood and therefore their use in causal inference can also avoid attenuation by errors (regression dilution bias).

This innovative utilization of SNPs as IVs rests on a triad of fundamental assumptions, which are indispensable for ensuring the validity of MR results (Burgess et al., 2017; Davey Smith and Ebrahim, 2003). Suppose now we want to make casual inference between the exposure X and the outcome Y using the instrument variable G. Conventional instrumental variable analysis requires that the instruments must meet three conditions:

A1 The IV G is associated with the exposure of interest X.

A2 G is independent of the confounding factors U that confound the association of X and the outcome Y.

A3 G is independent of outcome Y given X and the confounding factors U.

1.2 Methods

In this section, we first introduce several widely used methods based on the three core assumptions.

1.2.1 Individual-level Data MR

1.2.1.1 Wald method

The Wald estimator (Wald, 1940), or the ratio estimator is the simplest of estimating the causal effect of the exposure X on the outcome Y, and it uses a single instrumental variable G. If we regress X and Y separately on the IV G,

$$X = \beta_1 G + \varepsilon_1,$$

$$Y = \beta_2 G + \varepsilon_2,$$

and get the estimated value $\hat{\beta}_1, \hat{\beta}_2$, then the ratio estimate of the causal effect is

$$\hat{\beta} = \frac{\hat{\beta}_2}{\hat{\beta}_1} \tag{1.1}$$

Intuitively, we can think of the ratio method as saying that the change in the outcome Y caused by a unit increase in the exposure X is equal to the change in the X caused by a unit increase in the IV G, scaled by the change in the X caused by a unit increase in the G. To build a confidence interval for the estimate, we may use a normal approximation. The asymptotic variance of the ratio estimate (Thomas et al., 2007) is:

$$\hat{\sigma}_{\beta}^2 = \frac{var\left(\hat{\beta}_2\right)}{\hat{\beta}_1^2} + \frac{\hat{\beta}_2^2 var\left(\hat{\beta}_1\right)}{\hat{\beta}_1^4} - \frac{2\hat{\beta}_2 \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\hat{\beta}_1^3} \tag{1.2}$$

And the term $\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)$ will vanish if we estimate $\hat{\beta}_1$ and $\hat{\beta}_2$ from different samples. However, asymptotic normal approximations for the IV estimate may result in overly narrow confidence intervals, especially if the sample size is not large or the IV is weak. This is because IV estimates are not normally distributed. Alternatively, we can use Fieller's theorem (Fieller, 1954) or bootstrap method (Efron, 1992).

1.2.1.2 Two stage least square

Another popular used method in IV method is the two stage least square (2SLS or TSLS) regression. Suppose we have n individuals, and have the observed data $\{g_i, x_i, y_i; i = 1, ..., n\}$, $g_i \in \mathbb{R}^p, x_i \in \mathbb{R}, y_i \in \mathbb{R}$. Let$X = (x_1, ..., x_n)^T, Y = (y_1, ..., y_n)^T, G = (g_1, ..., g_n)^T \in \mathbb{R}^{n\times p}$.

In the first stage, the exposure X is regressed on the instrumental variables G to obtain the fitted values of X, denoted by $\hat{X}$:

$$X = G\beta_1 + \varepsilon_1, \qquad \hat{\beta}_1 = (G'G)^{-1}G'X, \tag{1.3}$$

$$\hat{X} = G\hat{\beta}_1 = HX, \qquad H = G(G'G)^{-1}G'. \tag{1.4}$$

In the second stage, the outcome Y is regressed on the fitted exposure $\hat{X}$:

$$Y = \hat{X}\beta_2 + \varepsilon_2, \tag{1.5}$$

$$\hat{\beta}_2 = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'HX)^{-1}X'HY. \tag{1.6}$$

Here, $\hat{\beta}_2$ estimates the effect of X on Y that is mediated solely through the component of X explained by the instruments G. Therefore, it is necessary to make sure the assumption (A1) is true. We can also build a confidence interval based on the variance of β_2.

$$var(\beta_2) = \sigma^2(X'HX)^{-1} \tag{1.7}$$

$$\hat{\sigma}^2 = (Y - \hat{X}\beta_2)'(Y - \hat{X}\beta_2)/(n-1)$$

1.2.1.3 Two sample two stage

Generally, it is hard to get a data set including all variables we need. On the contrast, collecting two separate samples, in which the first includes the IV G and exposure X and the second includes G and outcome Y, is much easier. In this situation, we can use a method called two sample two stage.

Suppose $G \in \mathbb{R}^{n\times p}$, $X \in \mathbb{R}^{n\times q}$. When $p = q$, we would have $G'X$ reversible. With exact identification, the causal effect we get in Section 1.2.1.2,

$$\hat{\beta}_{IV} = (G'X)^{-1}G'Y$$

Suppose now we only have two samples $\{Y_1, G_1\}$ and $\{X_2, G_2\}$, where $X_2 \in \mathbb{R}^{n_2 \times (k+p)}$ and $G_2 \in \mathbb{R}^{n_2 \times (k+q)}$. When $p = q$, Angrist and Krueger (Angrist and Krueger, 1999) proposed a consistent estimation of causal effect to use, which is

$$\hat{\beta}_{TSIV} = (G_2' X_2/n_2)^{-1}(G_1' Y_1/n_1) \tag{1.8}$$

Another statistics, valid also when $p \neq q$, named the two-sample two-stage least squares(TS2SLS) estimator is:

$$\hat{\beta}_{TS2SLS} = (G_2' X_2/n_2)^{-1} C (G_1' Y_1/n_1), \tag{1.9}$$

where $C = (G_2' G_2/n_2)(G_1' G_1/n_1)^{-1}$. Inoue and Solon (2010) have proved that $\hat{\beta}_{TS2SLS}$ is superior than $\hat{\beta}_{TSIV}$. Because the implicit correction for differences between the two samples in the distribution of G, matrix C, yields a gain in asymptotic efficiency.

The standard error of $\hat{\beta}_{TS2SLS}$ is

$$\hat{\varepsilon}_2 (\hat{X_2}' \hat{X_2})^{-1} \left(1 + \frac{n_1}{n_2} \frac{\hat{\beta}_{TS2SLS}' \hat{\Sigma}_{\varepsilon_1} \hat{\beta}_{TS2SLS}}{\hat{\varepsilon}_2} \right),$$

where $\hat{\varepsilon}_2$ is the sample mean squared residual from the second-stage regression, and $\hat{\Sigma}_{\varepsilon_1}$ is a consistent estimate of the covariance matrix for the first-stage disturbances.

This method can be naturally developed to be used when there is only summary data available. Further development of the method can be seen in Bowden et al. (2019), Zhao et al. (2020) and Minelli et al. (2021). It is also important to be aware of several potential limitations when using summary-level data, see Hartwig et al. (2021, 2016) for further discussion.

1.2.2 Summary-level Data MR

Individual level data on study participants are not always available due to issues of practicality and confidentiality of data-sharing. Burgess et al. (2013) outlines two approaches for estimating causal effects using summarized data. Assume that summary statistics are available for multiple genetic variants, each of which satisfies the IV assumptions. The models for the k-th variant are specified as follows:

$$X = \beta_{Xk} G_k + \varepsilon_{Xk}, \qquad \mathrm{Var}(\beta_{Xk}) = \sigma_{Xk}^2, \tag{1.10}$$

$$Y = \beta_{Yk} G_k + \varepsilon_{Yk}, \qquad \mathrm{Var}(\beta_{Yk}) = \sigma_{Yk}^2, \tag{1.11}$$

for $k = 1, \ldots, K$, where X_k, Y_k, σ_{Xk}^2, and σ_{Yk}^2 are assumed to be known.

1.2.2.1 Inverse-variance weighted combination of ratio estimates

For each genetic variant k, the ratio estimate of the casual effect of X on Y is β_{Yk}/β_{Xk}, and the standard error of the ratio estimate can be approximated using σ_{Yk}/β_{Xk}. The inverse-variance weighted (IVW) estimete of the causal effect combines the ratio estimates using each variant in a fixed-effect mata analysis model:

$$\hat{\beta}_{IVW} = \frac{\sum_k \beta_{Xk}\beta_{Yk}\sigma_{Yk}^{-2}}{\sum_k \beta_{Xk}^2 \sigma_{Yk}^{-2}} \tag{1.12}$$

$$se(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_k \beta_{Xk}^2 \sigma_{Yk}^{-2}}} \tag{1.13}$$

1.2.2.2 Likelihood-based method

We can also construct a model by assuming a linear relationship between the risk factor and outcome and a bivariate normal distribution for the genetic association estimates:

$$\begin{pmatrix} \beta_{xk} \\ \beta_{Yk} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \xi_k \\ \beta\xi_k \end{pmatrix}, \begin{pmatrix} \sigma_{Xk}^2 & \rho\sigma_{Xk}\sigma_{Yk} \\ \rho\sigma_{Xk}\sigma_{Yk} & \sigma_{Yk}^2 \end{pmatrix} \right) \tag{1.14}$$

Here the causal effect of X on Y is assumed to be the same β for all genetic variants k and can be estimated by direct maximization of the likelihood or by Bayesian methods.

Simulation results show that the power of estimates come from likelihood-based method is greater than that from the 2SLS method. And the inverse-level data analysis gives similar point estimates to an individual-level data analysis and slightly improved power over the likelihood-based method, but slightly too narrow confidence intervals.

However, there are always situations violating the three assumptions (A1)-(A3), then the methods above will no longer suitable and may produce large bias.

1.3 Key Methodological Challenges

Violations (Glymour et al., 2012) of the triadic assumptions underpinning MR analysis can produce biased and unreliable estimates. Specifically, the instrumental variables are called weak if they violate A1 (Staiger and Stock, 1994; Bound et al., 1995). And violation of A2 and A3 will lead to invalid instruments (Bowden et al., 2015; Kolesár et al., 2015; Hemani et al., 2018). In the following sections, we will review existing solutions that have been proposed to address these challenges.

1.3.1 Weak Instruments and Selection Bias

In MR analysis, two main frameworks are commonly used: two-sample MR analysis with GWAS summary statistics and one-sample MR analysis with individual-level data. While two-sample MR analysis has gained popularity due to easy access to public datasets, it comes with a couple of limitations. Firstly, it relies on marginal estimates of SNP statistics, which can be biased when not accounting for linkage disequilibrium (LD) properly. Secondly, it lacks the flexibility to model other causal mechanisms, such as nonlinear causal effects. As a result, there continues to be a significant interest in the advancement of statistical methods for one-sample MR analysis. The most popular method used in one-sample MR analysis is the two-stage least squares (2SLS) approach (Angrist and Krueger, 1991), which is relatively straightforward to implement and can yield consistent estimates of causal effects. However, the 2SLS estimate can be biased in the presence of weak instruments (Bound et al., 1995). The bias is in the direction of the confounded association and can cause inflated false positive rates, particularly when more than one IV is included in the analysis (Burgess et al., 2019). To date, weak instrument bias still remains one of the significant concerns in one-sample MR analysis (Burgess et al., 2019).

A potential solution to mitigate the impact of weak IVs is to opt for a two-sample MR analysis. While this approach might mitigate some biases, it does not eliminate them entirely. Specifically, bias due to weak instruments in two-sample MR tends to be directed towards the null (Angrist and Krueger, 1995). Limited information maximum likelihood (LIML) method (Anderson and Rubin, 1949; Anderson, 2005) was introduced as an alternative to 2SLS when dealing with weak instruments. Burgess et al. (Burgess et al., 2011) showed that LIML could provide a less biased estimate compared to 2SLS in the presence of weak instruments, but at the expense of incurring larger variance. Nevertheless, LIML is still subject to weak instrument problems and its finite sample performance can be poor. Angrist et al. (Angrist et al., 1999) proposed two jackknife instrumental variables estimators (JIVE) as alternatives to 2SLS and LIML to reduce the bias with many weak instruments. However, Sören and Matz (Blomquist and Dahlberg, 1999) showed that neither LIML nor the JIVE estimators perform uniformly better than the 2SLS does in terms of root mean square error.

In one-sample MR analysis, when the same dataset is used for both IV selection and causal effect estimation, the “winner’s curse" or IV selection bias emerges as another notable concern in addition to the weak IV bias issue (Burgess et al., 2019; Jiang et al., 2023). This bias could lead to biased causal effect estimates and hence inflate false positive rates under the 2SLS IV regression framework. This is evident in Appendix A.2.3, where it is shown that using the same data (the whole sample) for both IV selection and causal effect estimation, both LIML and 2SLS methods amplify bias compared to using half data for selection and the other half for causal effect estimation. Thus, it is critical to address the IV selection bias issue in one-sample MR analysis.

1.3.2 Pleiotropy and invalid IVs

Pleiotropy, the phenomenon where a single genetic variant influences multiple traits, is common in biology and genetics. In biological systems, pleiotropy is widespread and reflects the complex interplay between genes and traits. However, in the context of Mendelian Randomization, pleiotropy can undermine the validity of genetic variants used as IVs. Specifically, when a genetic variant affects the outcome not only through the exposure of interest but also via other independent pathways, it becomes an invalid IV and violates the core assumptions required for valid causal inference.

Pleiotropy in MR studies is typically categorized as either vertical or horizontal. Vertical pleiotropy occurs when a genetic variant influences an exposure, which in turn affects the outcome, and this aligns with the standard MR framework and does not violate IV assumptions. In contrast, horizontal pleiotropy arises when a genetic variant independently affects both the exposure and the outcome, leading to violations of the exclusion restriction assumption.

In practical MR analyses, identifying and removing invalid IVs is particularly challenging.

Researchers often rely on measures of instrument strength, such as p-values, to select relevant IVs. However, there is currently no statistically guaranteed procedure to reliably detect and exclude all invalid instruments. Moreover, when IV selection is performed using individual-level data from a single sample, selection bias may be introduced, further complicating causal effect estimation.

In recent years, many methods has been proposed to deal with invalid IVs. For summary statistics, under the InSIDE (Instrument Strength Independent of Direct Effect) assumption (Kolesár et al., 2015), Bowden et al. (Bowden et al., 2015) introduced MR-Egger regression, which identifies and corrects for horizontal pleiotropy using the intercept of a regression model. Then in 2016, under the majority rule, Bowden et al. (2016) proposed a weighted median estimator to provide consistent causal effect estimates even when up to 50% of the instrumental variables are invalid. Additionally, Verbanck et al. (2018) developed MR-PRESSO, which detects and corrects for horizontal pleiotropy by identifying and removing outlier instrumental variables and has been shown to perform best when horizontal pleiotropy affects less than 50% of the instruments. In 2021, Wang and Kang (2022) extended the Anderson-Rubin test (Anderson and Rubin, 1949), integrating the Kleibergen test (Kleibergen, 2002) and conditional likelihood ratio test to accommodate two-sample summary-data MR, improving robustness against weak and invalid IVs. Patel et al. (2024) introduced the Focused Instrument Selection method, which optimizes causal effect estimation by selecting invalid IVs with minimal direct effects under the local-to-zero assumption.

When we are available to individual-level data, there are some more methods can be used. In 2016, Kang et al. (2016) introduced the sisVIVE method, aimed at estimating causal effects without requiring complete knowledge of the validity of instrumental variables. The method is applicable under the majority rule and employs a penalized ℓ_1 estimation approach. However, despite its innovative framework, the method's accuracy in identifying valid instruments remains limited, often leading to estimates that still exhibit substantial bias. Furthermore, a key limitation of this method is its inability to perform inference as it provides an estimate of the causal effect but does not yield a standard deviation. In 2021, Windmeijer et al. (2021) proposed the CIIV method, which relaxes the assumptions on IVs required by sisVIVE. The CIIV method only requires the plurality rule, as introduced by Guo et al. (2018), which requires the valid IVs are the largest group having the same effects on the outcome. However, this method relies on a strong association between the IVs and the explanatory variables, meaning that when the IVs are weak, the accuracy of instrument selection still needs improvement. Apfel and Liang (2024) also proposed a method for selecting valid IVs using Agglomerative Hierarchical Clustering (AHC), which performs comparably to the CIIV in terms of selection and shows superior performance when dealing with multiple exposures. Ye et al. (2024) also proposed the GENIUS-MAWII method, which aims to provide robust Mendelian randomization inference in the presence of pervasive pleiotropy and a large number of weak instrumental variables. This approach leverages the heteroscedasticity of the exposure with respect to the instruments (and covariates) for identification; thus, if the required heteroscedasticity is absent, the method is not identifiable. Additionally, although GENIUS-MAWII can handle widespread pleiotropy, it relies on the key assumption that the effects of the instruments on the exposure and outcome do not interact with unmeasured confounders, which may not always hold in practice. Lin et al. (2024) proposed a method called WIT, which provides a detailed discussion on how to identify model parameters in the presence of weak IVs and employs the MCP penalty to select invalid IVs. Compared to previous methods, WIT significantly improves the accuracy of identifying invalid IVs. However, it still faces challenges in reliably constructing a trustworthy confidence interval. The estimates obtained using WIT are notably unstable, particularly in the presence of weak IVs, where numerous outliers significantly deviating from the true values may arise, as demonstrated in our subsequent simulations.

1.3.3 Nonlinearity and Multiple Exposures

What we have discussed so far are all assume the relationship between the exposure X and the outcome Y is linear. But sometimes observational data would suggest a non-linear association between the exposure X and the outcome Y, for example, alcohol consumption is consistently reported as having a U-shaped association with cardiovascular events (Marmot and Brunner, 1991). So it is necessary to extend MR methods to this kind of situation. Here we divide the nonlinear cases into three categories:

$$y = f(x)\beta + \varepsilon \tag{1.15}$$

$$y = x\beta(\theta) + \varepsilon \tag{1.16}$$

$$y = h(x, \beta) + \varepsilon \tag{1.17}$$

since it is hard to give a suitable explanation of the parameter β in reality in Model 1.16 and 1.17, what we consider here is majority the first situation, Model 1.15. Besides the methods what we list here, Burgess et al. (2014) also have provided an approach and done many applications in MR about the nonlinear exposure–outcome relationship. Singh et al. (2019) proposed kernel instrumental variable regression (KIV), a nonparametric generalization of 2SLS, and proved in experiments, KIV outperforms four kinds of methods for nonparametric IV regression.

1.3.3.1 Control function methods

The nonlinear of Model 1.15 is showed in the transformation of x. A classic method used in econometrics to address endogeneity is the control function approach (Wooldridge, 2015). This method is closely related to 2SLS and yields the same solution in linear models. But when the true model is nonlinear, the control function approach utilizes more information than 2SLS and can improve the precision of the estimates, albeit with some loss of robustness. For a detailed comparison of these two methods, see Guo and Small (2016). In addition, Sulc et al. (2022) conducted extensive simulations using the control function approach in Mendelian Randomization and demonstrated its strong performance.

For simplicity, we assume $f(x)$ in Model 1.15 is a polynomial function. Suppose the real model is

$$X = G\beta_1 + \varepsilon_x \tag{1.18}$$

$$Y = \sum_{j=0}^{k} \beta_{2j} X^j + \varepsilon_y \tag{1.19}$$

Since the error term ε_y and ε_x can be correlated due to confounders, we split ε_y:

$$\varepsilon_y = \sum_{j=0}^{l} \alpha_j \varepsilon_x^j + \tau_y \tag{1.20}$$

Then we have:

$$Y = \sum_{j=0}^{k} \beta_{2j} X^j + \sum_{j=0}^{l} \alpha_j \varepsilon_x^j + \tau_y = \sum_{j=0}^{k} \beta_{2j} X^j + \sum_{j=0}^{l} \alpha_j (X - G\beta_1)^j + \tau_y \tag{1.21}$$

So first we could regress X on G, to get the estimates $\hat{\varepsilon}_x$, then we regress Y on the transformation of X and $\hat{\varepsilon}_x$ to get the causal effects.

1.3.3.2 A more generalized method

A more generalized method for Model 1.15 is introduced by Li (2019). Unlike the typical assumptions applied in the linear setting, namely:

- **Relevance:** $\mathrm{Cov}(G, X \mid U) \neq 0$,
- **Exclusion restriction:** $\mathrm{Cov}(G, U) = 0$,

. Li proposes an alternative set of assumptions for the nonlinear model:

- **Relevance:**

$$\min_{f \in \mathcal{F}} \mathrm{loss}(X, f(G)) \leq \epsilon, \tag{1.22}$$

- **Exclusion restriction:**

$$\mathrm{loss}(V, v_\alpha(G)) \geq \mathrm{loss}(V, 0) - \epsilon', \tag{1.23}$$

where $V = Y - g_\beta(X)$, $g_\beta(X) \in \arg\min_{g \in \mathcal{G}} \mathrm{loss}(Y, g(X))$, and $v_\alpha(G) \in \arg\min_{v \in \mathcal{V}} \mathrm{loss}(V, v(G))$.

So in his methodology, stage one is to find:

$$\omega \in \arg\min_{\omega} \mathrm{loss}(X, f_\omega(G)) \tag{1.24}$$

This determines $\hat{X} = f_{\hat{\omega}}(G)$. Then in stage two, find

$$\beta \in \arg\min_{\beta} \mathrm{loss}(Y, g_\beta(\hat{X})), \tag{1.25}$$

such that $\mathrm{loss}(V, v_\alpha(G)) \geq \mathrm{loss}(V, 0) - \varepsilon$, where $V = Y - g_\beta(\hat{X})$.

However, his discussion is limited to the case where the nonlinear model is a generalized additive model (GAM), that is, $\hat{y} = b_0 + b_1 f_1(X) + \ldots + b_p f_p(X)$, where X denotes the input variable and y is the target variable. Therefore, future research could explore broader classes of nonlinear functions building upon his framework.

1.4 Structure of the Dissertation

In summary, we have reviewed the fundamental principles and core assumptions of MR methods, as well as the key challenges currently facing the field. In Chapter 2, we introduce the MR-SPLIT method, a framework based on 2SLS designed to address selection bias and the weak instrument problem in one-sample MR analyses. Building on this, Chapter 3 presents MR-SPLIT+, an enhanced approach that substantially reduces bias arising from invalid IVs and achieves significantly higher accuracy in identifying invalid IVs compared to existing methods. In Chapter 4, we further extend MR-SPLIT+ to accommodate more complex scenarios, proposing the BiMR-SPLIT+ method for bidirectional MR studies, which offers additional improvements over MR-SPLIT+. Finally, Chapter 5 summarizes our main contributions and discusses potential directions for future research.

CHAPTER 2

MR-SPLIT - ADDRESSING SELECTION AND WEAK INSTRUMENT BIAS

2.1 Introduction

In one-sample MR analysis, IVs are typically chosen based on a p-value threshold. However, the usage of a p-value threshold criterion in the selection of IVs is somewhat arbitrary and lacks robust justification. The 2SLS approach relies on the fitted values from the first stage for estimating causal effects in the second stage, highlighting the critical role of prediction accuracy and thus questioning the robustness of models that depend solely on p-value thresholds for validation. Given the typically vast dimensionality of SNP data, the use of penalized shrinkage methods can effectively mitigate the winner's curse effect in one-sample MR analysis. This strategy prioritizes prediction accuracy and hence provides a potentially more dependable and robust framework for causal inference.

Denault et al. (Denault et al., 2022) introduced a method called 'Cross-Fitting for Mendelian Randomization' (CFMR) to handle the weak instrument issue in one-sample MR analysis, which consolidates information from multiple IVs into a single IV, termed the Cross-Fitted Instrument (CFI). CFMR randomly splits a sample into K subgroups $\{I_1, \cdots, I_K\}$ and define the complement of the partition I_k as $I_k^c = \{1, \cdots, N \notin I_k\}$. In each subset $\{I_k^c\}$, it first selects γ_k independent variants $\{Z_{1,k}, \cdots, Z_{\gamma_k,k}\}$, and then defines a CFI of the exposure X on I_k, which is the prediction of X on I_k trained using data with indexes in I_k^c.

This predicted value can be viewed as a polygenic risk score in risk prediction analysis. Then, the new CFI is used as the IV to fit the 2SLS model for further causal inference. CFMR consolidates all the IVs into one single IV (CFI), thus it produces less biased results. However, CFMR does not completely solve the selection bias issue. Taking $K = 10$ as an example, CFMR employs 9 folds of data for selecting IVs and applies the estimated effects to construct the composite IV in a separate fold of data. By iterating this process 10 times, the composite IVs across any pair of folds are constructed with 80% of data in common. Thus, the composite IVs are not constructed using completely independent data. This could lead to a new manifestation of the winner's curse problem. Furthermore, by relying on one CFI as the only IV to represent the collective information of all IVs, there could be potential information loss which further leads to variance inflation and consequently reduced power (as shown in our theorem and simulation studies).

In general, the selection of IVs involves a bias and variance trade-off when estimating the causal effect. Using more IVs tends to introduce a larger bias but smaller variance, whereas employing too few IVs results in a smaller bias but larger variance. Pierce et al. (Pierce et al., 2010) did intensive simulations to evaluate the power and IV strength requirements for MR analyses based on 2SLS. They employed four strategies to combine information across IVs and evaluated the consequences of these strategies on power and overall IV strength, as measured by the first-stage F statistic in 2SLS. The results suggest that categorizing IVs into major and weak ones and then consolidating the weak ones into a single IV based on the knowledge of the genetic architecture underlying the exposure, can mitigate the issue of weak IVs. However, the study identifies a gap in current methodologies: it does not provide a clear approach for differentiating between major and weak IVs, nor does it offer a strategy for combining weak IVs in the context of one-sample MR analysis. This highlights an area for further research and methodology development in the field.

In this chapter, we propose an adaptive Sample-sPLitting method with cross-fitting InstrumenTs (MR-SPLIT) to address the bias issue of IV selection and weak instruments. This approach can effectively reduce the number of weak IVs without the loss of much information, thereby enhancing the performance of causal inference in MR studies by improving the power of causal inference. Our method has two advantages over the existing ones: 1) It adaptively selects major and weak IVs, subsequently creating a composite IV from the weaker ones. We theoretically proved that the variance of the MR-SPLIT estimate is smaller than that of the CFMR estimate under the condition of one sample split. Simulation results also show that MR-SPLIT can always achieve higher power and lower RMSE than CFMR; and 2) A multi-sample splitting strategy is further employed to enhance the robustness of estimation and testing. Extensive simulation studies were conducted to assess the performance of our method in comparison to its counterparts, including 2SLS, LIML, and CFMR. Our method offers an efficient and powerful solution for one-sample MR analysis by addressing two primary sources of bias: IV selection bias and the bias associated with weak instruments.

2.2 Statistical Method

Assume the following structural equation model,

$$y_i = x_i\beta + \varepsilon_{yi}$$

$$x_i = G_{i\cdot}\alpha + \varepsilon_{xi}$$

where x_i is the exposure, and y_i denotes the outcome of the ith individual. $G_{i\cdot}$ is a p-dim vector of SNP IVs, where $G_{i\cdot} = \{G_{i1}, G_{i1}, \dots, G_{ip}\} \in \mathbb{R}^p$. The error term is denoted by $\varepsilon_i = (\varepsilon_{xi}, \varepsilon_{yi}) \sim N(0, \sigma^2 R)$ where $R_{12} (= \rho)$ is the correlation due to confounding. β is the interested causal effect which needs to be estimated. Suppose we have N independent individuals, and denote $Y = (y_1, \dots, y_N)' \in \mathbb{R}^{N\times 1}$, $X = (x_1, \dots, x_N)' \in \mathbb{R}^{N\times 1}$, $G = \{G_1, \dots, G_p\} \in \mathbb{R}^{N\times p}$, where the jth IV denoted as $G_j = (G_{1j}, \dots, G_{Nj})'$, $j = 1, \dots, p$, then we have

$$\begin{aligned} Y &= X\beta + \varepsilon_y \\ X &= G\alpha + \varepsilon_x \end{aligned} \tag{2.1}$$

2.2.1 Cross-fitting Instruments with Sample Split

Given the observed data $\{X, Y, G\}$, we first need to select a valid IV subset from the existing SNP pool, where the number of SNPs can be much larger than the sample size. To reduce potential biases and enhance the accuracy of estimates in MR analysis, one can use one sample for the selection of appropriate IVs and a separate, independent sample for the 2SLS estimation. By doing so, over-fitting and biases stemming from sample-specific peculiarities, such as the double dipping issue, can be minimized, leading to more robust and credible causal effect estimates. When only one sample is available, one simple idea is to randomly split the data into two equal subsets $\{I_1, I_2\}$, each containing roughly $N/2$ samples. Then, one can use one subset (say I_1) to select the IVs and use the other (say $I_2 = I_1^c$) to get the estimates of β.

For the IV selection, if no prior information about specific SNPs is available, researchers usually regress the exposure variable on each SNP, and then select those SNPs that yield marginal p-values smaller than a preset threshold (e.g., 5×10^{-8}) followed by LD pruning or LD clumping.

However, such a threshold is quite ad hoc and sometimes can be too stringent, prompting the need for relaxation, as advocated in some studies (Panagiotou et al., 2011). Such strictness can lead to the exclusion of valid IVs and the loss of valuable information. Conversely, if the threshold is too lenient, it may result in the selection of an excessive number of SNP IVs, potentially introducing challenges associated with weak IVs (Burgess et al., 2011). We suggest using some high-dimensional screening methods such as sure independence screening (SIS) (Fan and Lv, 2008) to first reduce the SNP dimension from ultra-high to high dimension. Methods like SIS have the sure screening property in which they ensure that, as the sample size increases, the probability of including all relevant variables becomes close to one. After this step, shrinkage methods such as LASSO or adaptive LASSO (Tibshirani, 1996; Zou, 2006) can be employed to select and estimate non-zero SNP effects. Other penalized methods with different penalty functions such as MCP or SCAD can also be applied.

After the SNP selection, directly employing these IVs in 2SLS might lead to the issue of weak instruments, potentially resulting in biased estimate. To mitigate this, we group the IVs into two groups, major IVs and weak IVs, based on their association strength with the exposure. Conventionally, the validity of IVs is assessed using F statistics. A common benchmark used in econometrics and statistical literature suggests that an F-statistic exceeding 10 is indicative of strong instruments, particularly when assessing the strength of a collective set of IVs (Stock and Yogo, 2002; Shea, 1997). However, the determination of the weakness of an individual IV lacks a widely recognized standard. In this analysis, we employed partial F-statistics with different thresholds as criteria for selecting major IVs. Generally, the partial F statistic is defined as:

$$F = \frac{\left(\mathrm{RSS}_r - \mathrm{RSS}_f\right)/p}{\left(\mathrm{RSS}_f\right)/(N-k-1)}$$

where RSS_r and RSS_f are the residual sums of squares for the reduced and full model, respectively; N is the total number of observations; k and p are the numbers of variables in the reduced and full model, respectively. This statistic measures how much the addition of p variables improves the model, compared to the increase in complexity these variables bring. It is a good statistic for calculating the strength of each IV and is consistent with the commonly used F-statistic for evaluating IV strength. In our model, $p = 1$ because we calculate the partial F statistics for each IV. The thresholds were set at partial F-statistics greater than 10, 30, and 50. We conducted a simulation study to compare these three statistics for the purpose of identifying weak IVs, as detailed in section 2.3.1. Based on the simulation results, it is recommended to use a threshold of partial $F > 30$ to define major IVs.

2.2.2 Composite IV for Weak Instruments

Following the separation of major and weak IVs, we propose consolidating the weak ones into a composite IV. Then, the major IVs and the composite IV are included in the 2SLS model to infer the causal effect (see Fig 2.1 for the flowchart of MR-SPLIT). By only consolidating the weak IVs into a single instrument, we can substantially reduce the number of IVs in the model while retaining most of the information they carry.

Denote the selected index of weak IVs as $S_{k,W}, k = 1, 2$, where $|S_{k,W}| = p_2$ represents the selected numbers of weak IV using data in I_k. Taking sample I_1 as an example, let the estimated effects for the weak IVs on the exposure be denoted as $\hat{\alpha}_{1,W} = \{\hat{\alpha}_{1,j}; j \in S_{1,W}\}$ for data in I_1. Here, the subscript 1 indicates that this parameter is estimated from subsample I_1, and the subscript W signifies that it corresponds to the direct effect of weak IVs on X from Eq (2.1). The new composite IV constructed in sample I_2, $\hat{G}_{2,W}$, is then defined as

$$\hat{G}_{2,W} = \sum_{j \in S_{1,W}} \omega_j G_{2,j} \tag{2.2}$$

where

$$\omega_j = sign(\hat{\alpha}_{1,j}) \frac{|\hat{\alpha}_{1,j}|}{\sum_{j \in S_{1,W}} |\hat{\alpha}_{1,j}|} \tag{2.3}$$

In other word, we use weak IVs selected from sample I_1 to construct the new composite IV in sample I_2. Then, we can use the major IVs and the new composite IV, i.e., $\{G_{2,M}, \hat{G}_{2,W}\}$, to get the cross-fitted exposure in I_2. Here, the subscript M represents that $G_{2,M}$ is identified as the major IV, while the subscript W signifies that $\hat{G}_{2,W}$ is estimated from the weak IV. The subscript 2 indicates that these values are obtained from subsample I_2.

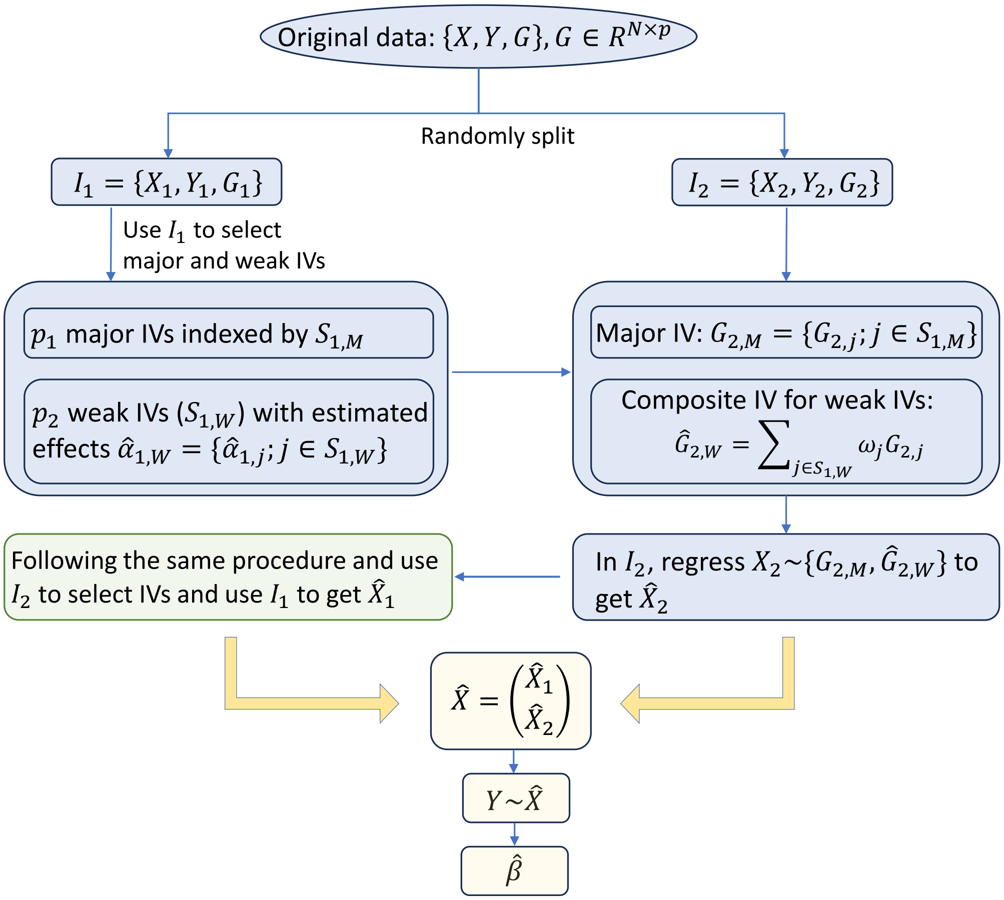

Original data: $\{X, Y, G\}, G \in R^{N\times p}$
Randomly split
$I_1 = \{X_1, Y_1, G_1\}$
$I_2 = \{X_2, Y_2, G_2\}$
Use I_1 to select major and weak IVs
p_1 major IVs indexed by $S_{1,M}$
p_2 weak IVs ($S_{1,W}$) with estimated effects $\hat{\alpha}_{1,W} = \{\hat{\alpha}_{1,j}; j \in S_{1,W}\}$
Major IV: $G_{2,M} = \{G_{2,j}; j \in S_{1,M}\}$
Composite IV for weak IVs: $\hat{G}_{2,W} = \sum_{j\in S_{1,W}} \omega_j G_{2,j}$
In I_2, regress $X_2 \sim \{G_{2,M}, \hat{G}_{2,W}\}$ to get $\hat{X}_2$
Following the same procedure and use I_2 to select IVs and use I_1 to get $\hat{X}_1$
$\hat{X} = \begin{pmatrix}\hat{X}_1\\ \hat{X}_2\end{pmatrix}$
$Y \sim \hat{X}$
$\hat{\beta}$

Figure 2.1 The flow chart of MR-SPLIT with one random split.

Note: The original data is randomly split into two parts indexed by I_1 and I_2. We use data in I_1 to select major and weak IVs, then form the composite IV for the weak ones in I_2 with the weight ω being calculated based on Eq (2.3), then get the fitted $\hat{X}_2$ in I_2. Similarly, we use data in I_2 to select major and weak IVs, then use data in I_1 to form the composite IV and get the fitted values $\hat{X}_1$. Next, we combine $\hat{X}_1$ and $\hat{X}_2$ to get $\hat{X} = (\hat{X}_1^T, \hat{X}_2^T)^T$ and fit the second stage regression model $Y \sim \hat{X}$ to get the causal estimate ($\hat{\beta}$) and its p-value.

To clarify logic, for each $k = 1, 2$ we use the subset I_k to identify the SNP IVs and obtain the estimated effects $\hat{\alpha}$ for the selected IVs. Then, we categorize them into two groups, major IVs and weak IVs, using the partial F-statistic criterion defined earlier. We then combine the weak IVs in I_k^c using the estimated weights from I_k. This approach enables us to avoid overfitting by selecting IVs and estimating the causal effect using different samples.

2.2.3 Estimating the causal effect

Once we get the IVs $\{G_M, \hat{G}_W\}$ in each $I_k, k = 1, 2$, we can then perform the first stage of the 2SLS regression on these IVs to get the cross-fitted exposures $\hat{X}_k$ which are then aggregated, i.e.,

$$\hat{X} = \begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \end{pmatrix} \in \mathbb{R}^{N\times 1}.$$

The causal effect can be estimated by regressing Y on $\hat{X}$ using the whole sample, which is given by

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y \tag{2.4}$$

Cross-fitting allows for the utilization of the entire dataset in estimating causal effects, thereby circumventing the winner's curse problem that arises when the same data is employed for both IV selection and causal effect estimation.

Remark 1: Both MR-SPLIT and CFMR implement a cross-fitting idea with sample splitting, but the analysis is fundamentally different. MR-SPLIT combines the cross-fitted exposures for further causal inference, while CFMR combines cross-fitting instruments for further causal inference. CFMR first calculates the composite IV in the kth split sample (denoted as $\tilde{G}_{k,n_k\times 1}$), where n_k denotes the sample size in the kth split sample, then combines these composite IVs to form the final composite IV (denoted as $\tilde{G}_{N\times 1}$ by stacking all $\tilde{G}_k$), and finally uses the full data $(X, \tilde{G}, Y)$ to perform 2SLS analysis for causal inference. Each composite IV $\tilde{G}_{k,n_k\times 1}$ can be regarded as a polygenic risk score based on the kth split sample. As the dimensions of major and weak IVs identified from different sample splits are different, CFMR is infeasible to separate the two components and incorporate them in downstream causal inference.

Remark 2: CFMR uses data in I_k^c to select IVs, then calculates the composite IV based on data in I_k. Typically, a 10-fold split suffices. On the other hand, MR-SPLIT benefits from a 2-fold sample splitting. It uses data in I_1 to select and separate major and weak IVs, then forms the cross-fitting composite IV in I_2. After that, it calculates the cross-fitted exposures based on the major IV(s) and the cross-fitting composite IV for further causal inference. More data for IV selection leads to less data to fit the cross-fitted exposure and vice versa. To balance the two components, a 2-fold sample split is recommended.

Remark 3: By combining the cross-fitted exposures, Theorem 1 shows that MR-SPLIT produces estimates with a variance no larger than that of CFMR if both approaches implement a 2-fold sample split. The proof is given in This finding extends to scenarios involving a $k(> 2)$-fold sample split for CFMR. Although providing theoretical proof for this result poses a challenge, we have demonstrated its validity through simulations.

Theorem 1. *Let $\hat{\beta}_{CFMR}$ and $\hat{\beta}_{MR-SPLIT}$ be the 2SLS estimates obtained respectively by the CFMR and MR-SPLIT method with a 2-fold random split. Then, $\hat{\beta}_{MR-SPLIT}$ is more efficient than $\hat{\beta}_{CFMR}$ in the sense that*

$$var(\hat{\beta}_{MR-SPLIT}) \le var(\hat{\beta}_{CFMR})$$

The proof of Theorem 1 is given in Appendix A.2.

2.2.4 Multiple Sample Splitting and Robustness

Given the inherent uncertainty in single-sample splitting, particularly in cases of limited sample size, we propose a multiple-splitting strategy to improve the robustness of the approach. We randomly split data (into two halves) L times. For each random split, the same estimation and testing procedure as described before are conducted. Let $pval_l$ denote the p-value at the lth random split. There are different ways to aggregate these L p-values. One approach involves employing the aggregation method for p-values proposed by Wasserman and Roeder (Wasserman and Roeder, 2009; Dezeure et al., 2015). However, this method has proven to be overly conservative in our simulations. Another simple way is to use the Cauchy combination rule for correlated p-values (Liu and Xie, 2019), which is similar to the minimum p-value method but does not require an intensive resampling procedure to assess the null distribution of the minimum p-value. Given its computational efficiency, we adopt the Cauchy combination rule to aggregate p-values obtained from multiple sample splitting. Following (Liu and Xie, 2019), the test statistics is defined as:

$$T_{cauchy} = \sum_{l=1}^{L} \omega_l \tan\left((0.5 - pval_l)\pi\right)$$

where the weights ω_l are non-negative and $\sum_{l=1}^{L} \omega_l = 1$. If no further information is available, the weight ω_l can be simply chosen as $1/L$. The p-value of T_{cauchy} can be simply approximated by

$$\text{p-value} = \frac{1}{2} - \arctan(T_{cauchy})/\pi \tag{2.5}$$

In essence, augmenting the number of sample splits improves result robustness. However, this enhancement comes with the trade-off of requiring increased computational resources. To provide general guidance on the number of splitting times, we conducted a simulation study (see section 2.3.4). The results suggest that conducting multiple splits about 50-60 times is sufficient to achieve a robust outcome in terms of controlling type I errors and maintaining stable statistical power. In the case of a large sample size and strong SNP heritability, the splitting time can be dramatically reduced (see the simulation results).

2.2.5 Algorithmic Details

The detailed algorithm of the MR-SPLIT is given below:

1. For each $l = 1, \cdots, L$ random split, repeat the following steps:

    a) Split the sample into two equal subsets $\{I_1, I_2\}$, i.e., $\{1, \cdots, N\} = I_1 \cup I_2$ with $I_1 \cap I_2 = \emptyset$ and $|I_1| = [N/2]$ and $|I_2| = N - [N/2]$, and denote the complementary sets as $\{I_1^c, I_2^c\}$ accordingly.

    b) For each $k = 1, 2$, we use I_k^c to select IVs and get the estimated effect size for each IV. Then, categorize the selected IVs into two distinct groups, major IV(s) and weak IVs, based on the partial $F > 30$ criterion.

    c) In each subset I_k, combine the weak IVs using the effect size estimated from I_k^c following Eq (2.2). Then regress the exposure variable X on the new IVs (major IV(s) + composite IV) to get the fitted value $\hat{X}_k$.

    d) Denote $\hat{X} = \begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \end{pmatrix}$, and do the second stage regression of Y on $\hat{X}$ to get the causal effect estimate $\hat{\beta}_l$ and the p-value $pval_l$.

2. Calculate the Cauchy combination statistics $T_{\text{cauchy}} = \frac{1}{L}\sum_{l=1}^{L}\tan((0.5 - pval_l)\pi)$, and the aggregated p-value as $pval = \frac{1}{2} - \arctan(T_{cauchy})/\pi$. The final aggregated causal effect estimate can be calculated as $\hat{\beta} = \frac{1}{L}\sum_{l=1}^{L}\hat{\beta}_l$.

2.3 Simulation Study

We conducted simulations to assess the performance of our method and provided guidance on the identification of major IVs and selecting an efficient number of sample splitting. Subsequently, we compared the proposed MR-SPLIT with the existing approaches, including 2SLS, LIML, and CFMR, across various settings.

2.3.1 Major IV Identification

We applied 3 criteria, $F > 10$, $F > 30$, and $F > 50$, to distinguish the major and weak IVs under various settings. We randomly generated 300 independent SNPs each with MAF=0.3, and assumed only 5 SNP had effects on the exposure. The effects of these SNPs were set to be $\beta = (0.4, 0.4, 0.1, 0.05, 0.05)\omega_0$, where ω_0 was chosen to ensure that these SNPs account for $h^2 = \{0.15, 0.30, 0.50\}$ of the variation in exposure (h^2 can be interpreted as the exposure heritability). The error term was assumed to follow the standard normal distribution with mean 0 and variance 1. The rest 295 SNPs were assumed to be noises with no effect on the exposure (i.e., $\beta = 0$). Then, we followed model (2.1) to simulate the exposure. In this setting, the initial two SNPs may be regarded as the major IVs, whereas the remaining three are categorized as weaker ones. However, this differentiation can also be contingent on the signal-to-noise ratio, meaning that the first two SNPs may not be deemed as the major ones when h^2 is low, say $h^2 = 0.15$. And when the IVs are strong enough, say $h^2 = 0.5$, the three weaker IVs may be regarded as strong IVs. After applying SIS screening and LASSO estimation on these 300 SNPs, we then used these three criteria to distinguish the major and weak IVs. Part of the results can be seen in Table 2.1. As we mentioned before, there were 295 noise SNPs in total. It is possible that some of these noise SNPs may be incorrectly identified as major IVs. We also summarized these results in the last column of Table 2.1. More detailed information can be found in Table A.1 and Fig A.1 in Appendix A.2. Our analysis indicates that employing a partial $F > 10$ threshold to define major

IVs is excessively lenient, leading to misidentifying noises as major IVs, particularly in scenarios with small sample sizes (e.g., $N = 500$). Conversely, a threshold of $F > 50$ proves overly stringent, failing to recognize SNP 1 and 2 as major IVs in conditions characterized by low sample sizes and heritability. A threshold of $F > 30$ emerges as a balanced criterion for defining major IVs, effectively mitigating the aforementioned issues. Thus, we propose to use a partial $F > 30$ threshold in the selection of major IVs.

Table 2.1 Mean numbers of being identified as major IV using different criteria in 1,000 simulations.

h^2	N	Criteria	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	Noises (×295)*
0.15	500	F>10	0.55	0.5	0.15	0	0.1	1.25
		F>30	0.05	0.05	0	0	0	0
		F>50	0	0	0	0	0	0
	1000	F>10	0.95	0.95	0.35	0.1	0	0.65
		F>30	0.35	0.35	0	0	0	0
		F>50	0.15	0	0	0	0	0
	2000	F>10	1	1	0.75	0.35	0.55	0.6
		F>30	1	0.95	0.1	0	0	0
		F>50	0.75	0.75	0	0	0	0
0.3	500	F>10	0.95	0.8	0.25	0.25	0.1	1.15
		F>30	0.5	0.55	0	0	0	0
		F>50	0.25	0.25	0	0	0	0
	1000	F>10	1	1	0.75	0.45	0.3	0.65
		F>30	1	1	0.2	0	0	0
		F>50	0.8	0.7	0.05	0	0	0
	2000	F>10	1	1	1	0.75	0.9	0.6
		F>30	1	1	0.6	0.15	0.1	0
		F>50	1	1	0.1	0	0.05	0
0.5	500	F>10	1	1	0.8	0.35	0.4	1.8
		F>30	1	1	0.35	0	0.05	0
		F>50	0.9	0.9	0	0	0	0
	1000	F>10	1	1	1	1	0.95	0.6
		F>30	1	1	0.8	0.5	0.15	0
		F>50	1	1	0.4	0.05	0	0
	2000	F>10	1	1	1	1	1	0.4
		F>30	1	1	1	0.9	0.9	0
		F>50	1	1	1	0.4	0.3	0

*The total number of noise SNPs incorrectly identified as major IVs out of the 295 noise SNPs.

2.3.2 Comparison with 2SLS and LIML

We compared the proposed MR-SPLIT with the widely-used 2SLS approach and the LIML method which is particularly designed to address the weak instruments bias issue. We simulated 300 SNPs independently and randomly selected 5 SNPs as the IVs to generate the exposure variable X. We set $h^2 = \{0.15, 0.3, 0.5\}$ which respectively represent weak, moderate, and strong overall effect, and $\rho = (0.1, 0.2)$ where $\rho = \mathrm{cor}(\varepsilon_{xi}, \varepsilon_{yi})$ controls the unknown confounding effect.

We set the sample size (N) to 1000. To ensure a fair comparison with 2SLS and LIML, we only split the sample once (i.e., no multiple splitting). We then used one subset for selecting the IVs and incorporated the other subset with the selected IVs for estimation. Both 2SLS and LIML followed the same process but did not differentiate between major and weak IVs for further causal inference. To check the impact of selection bias for 2SLS and LIML, we also did the analysis using the whole dataset for both IV selection and causal effect estimation. The simulation settings are the same as what we previously described. The only difference is that we do not split the sample and use the whole sample to do the IV selection and estimation. Results for this analysis were given in Appendix A.2. The respective boxplots, illustrating the distribution of estimations across 1000 simulation iterations, are provided in Figs A.2, A.3 and A.4 in Appendix A.2. It is evident from the results that using the entire sample for both IV selection and effect estimation results in estimates with smaller variance but larger bias, leading to a significantly higher type I error rate. In the following, we only show the results based on sample splitting.

Table 2.2 presents a comparative analysis of the estimation accuracy among MR-SPLIT, LIML, and 2SLS. It shows that MR-SPLIT can provide estimates with a significantly small bias. In contrast, the estimates from 2SLS exhibit large bias, especially under weak IV and substantial confounding effects (e.g., $\rho = 0.2$). In some of the cases, LIML gives a smaller bias than MR-SPLIT does, but it has consistently larger variance than MR-SPLIT, leading to a conservative coverage probability (CP) compared to MR-SPLIT. The variance of 2SLS is uniformly smaller than the other two methods. However, given its large bias, it has the most poor coverage probability among the three methods. On the other hand, MR-SPLIT shows consistently good coverage probabilities under different scenarios, showcasing its robust performance under different conditions.

Table 2.2 Simulation comparison between M* (MR-SPLIT), LIML and 2SLS.

h^2	ρ	β	Bias($\lvert\beta - \hat{\beta}\rvert \times 100$)			Est. SE			CP*		
			M*	LIML	2SLS	M*	LIML	2SLS	M*	LIML	2SLS
0.15	0.1	-0.08	0.18	0.17	4.86	0.1252	0.1776	0.0788	0.955	0.827	0.895
		0.08	0.82	0.23	5.05	0.1253	0.1884	0.0795	0.957	0.832	0.886
	0.2	-0.08	0.17	1.19	9.45	0.1230	0.1737	0.0782	0.949	0.843	0.740
		0.08	0.31	1.01	9.44	0.1320	0.1770	0.0801	0.952	0.825	0.732
0.30	0.1	-0.08	0.02	0.11	2.4	0.0520	0.0844	0.0605	0.958	0.895	0.929
		0.08	0.13	0.59	2.94	0.0515	0.0831	0.0600	0.959	0.898	0.919
	0.2	-0.08	0.43	0.67	4.77	0.0501	0.0840	0.0612	0.946	0.904	0.837
		0.08	0.47	0.31	5.03	0.0524	0.0840	0.0621	0.947	0.905	0.845
0.50	0.1	-0.08	0.32	0.08	1.12	0.0329	0.0482	0.0430	0.938	0.943	0.945
		0.08	0.11	0.22	0.82	0.0328	0.0474	0.0424	0.948	0.931	0.944
	0.2	-0.08	0.14	0.02	2.08	0.0335	0.0513	0.0457	0.942	0.909	0.902
		0.08	0.2	0.03	2.07	0.0318	0.0469	0.0423	0.954	0.938	0.924

CP*=coverage probability

Fig 2.2 shows the results of the type I error of the three methods. We can observe that MR-SPLIT can effectively control Type I errors, even in the presence of strong unknown confounding. As depicted in Fig 2.2, both LIML and 2SLS methods exhibit much poorer performance than MR-SPLIT. Notably, 2SLS suffers from poor type I error control when the confounding effect is strong (i.e., $\rho = 0.2$), leading to inflated error rates. LIML has high false positive rates when the SNP effects are weak (i.e., weak instruments with low h^2), especially under $\rho = 0.2$. As the SNP effects increase, its performance improves; however, it can only effectively control type I errors when the instrumental variables are strong, as demonstrated in the scenario with $h^2 = 0.5$. Conversely, MR-SPLIT consistently demonstrates robust type I error control under all conditions, even under $h^2 = 0.15$ and $\rho = 0.2$, where 2SLS and LIML exhibit their poorest performance. The inflated type I error rates lead to inflated statistical power for 2SLS and LIML. Consequently, comparing power between MR-SPLIT and these two methods may not be a fair comparison; thus, we did not show the detailed power comparison here. Nevertheless, in the scenario where $h^2 = 0.5$ and $\rho = 0.1$, MR-SPLIT still attains the highest power, reaching 0.683 compared to 0.545 for 2SLS and 0.419 for LIML.

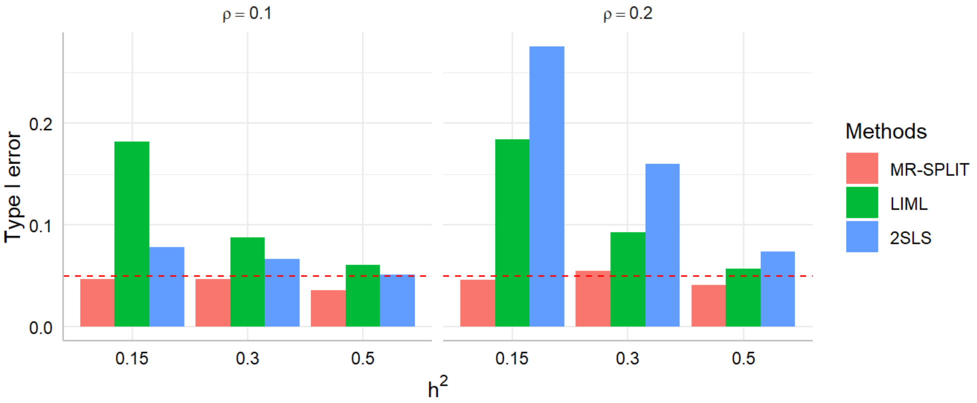


Figure 2.2 Type I error comparison between MR-SPLIT, 2SLS and LIML. The horizontal dashed line denotes the 0.05 level.

2.3.3 Comparison with CFMR

We compared our method with CFMR under different simulation scenarios. To ensure a fair comparison with CFMR, we applied 10-fold CFMR as recommended in the CFMR work, and 2-fold MR-SPLIT with 50 random sample splits. We applied the same procedure for selecting IVs. While CFMR combined all the selected IVs into a single composite one, our method differentiated between major and weak IVs using the partial $F > 30$ criterion and only weak IVs were combined into a composite one. We also followed the simulation settings described in the CFMR work to ensure a fair comparison. We generated a set of 300 SNPs, and the minor allele frequency is fixed as 0.3 for all the SNPs. We randomly chose 5 SNP IVs to generate the exposure variable with the model $X = \sum_{j=1}^{5} G_j\alpha_j + \varepsilon_x$, and the outcome with the model $Y = X\beta + \varepsilon_y$, where

$$\begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \sim N\left(0,\ 5\begin{pmatrix} 1 & 0.16 \\ 0.16 & 1 \end{pmatrix}\right)$$

We set two scenarios to comprehensively compare MR-SPLIT and CFMR:

- Scenario I: The effect sizes of the 5 SNPs are different, i.e., $\alpha = (0.4, 0.4, 0.1, 0.05, 0.05)$. Potentially, SNPs with the effect of 0.4 can be regarded as major IVs and the rest can be considered as weak ones. This also depends on the SNP heritability level h^2.

- Scenario II: The effect sizes of the 5 SNPs are the same, i.e., $\alpha = (0.2, 0.2, 0.2, 0.2, 0.2)$. In this case, differentiating between major and weak IVs can be challenging, presenting a less favorable condition for our method.

In each scenario, we compared the two methods in different aspects by changing the sample size ($N = \{1000, 3000, 5000\}$), variation in the exposure explained by the SNP IVs ($h^2 = \{0.15, 0.2, 0.3\}$) and the exposure's effect size for β.

Fig 2.3 depicts the type I error control of the two methods in scenario I and scenario II. In general, the control of type I error for the two methods is highly comparable across different settings characterized by distinct sample sizes and SNP heritability levels. Though the type I error is a little inflated for MR-SPLIT under a small sample size ($N = 1000$), particularly in scenario II, it controls the type I error well as the sample size increases.

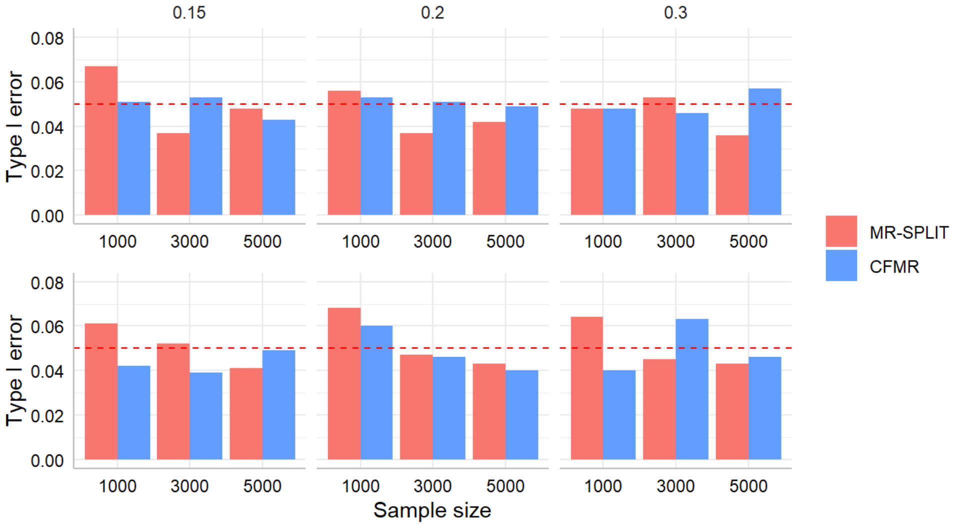


Figure 2.3 Comparison of type I error between MR-SPLIT and CFMR in Scenario I (top) and II (bottom). The horizontal dashed line denotes the 0.05 level.

Fig 2.4 shows the results of the power for the two methods in these two scenarios. Regardless of the settings, MR-SPLIT consistently exhibits higher power than CFMR. This discrepancy becomes especially noticeable when the IVs are relatively weak (i.e., $h^2 = 0.15$).

In Appendix A.2, we also presented the estimation performance of both methods when $\beta = 0$ and 0.08 in Figs A.5-A.10. The results reveal minimal difference in the causal effect estimation

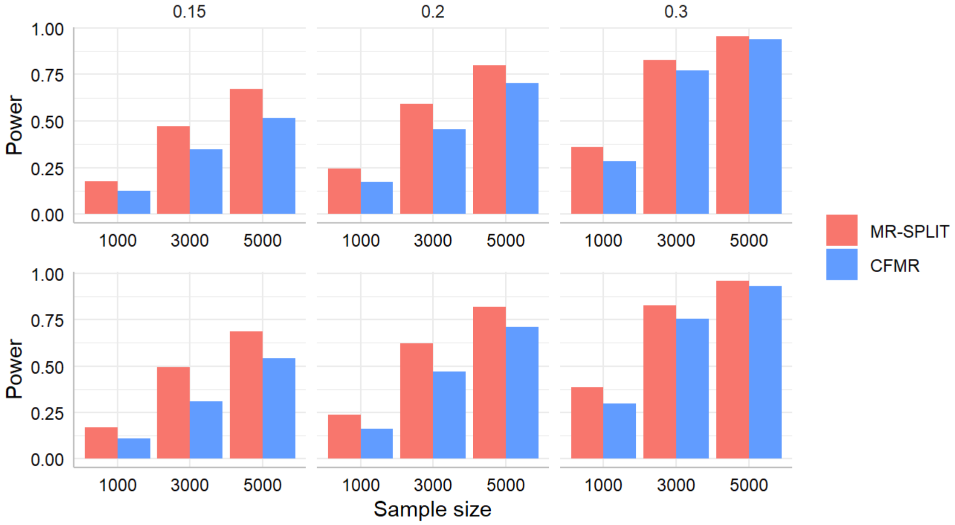


Figure 2.4 Power comparison between MR-SPLIT and CFMR in Scenario I (top) and II (bottom).

between the two methods. However, a noticeable distinction is the smaller standard error observed in MR-SPLIT across nearly all the scenarios, resulting in a smaller Root Mean Square Error (RMSE) (see Fig A.11 in Appendix A.2) and higher statistical power when compared to CFMR. This aligns well with the theoretical finding in Theorem 1, though the result is proved under a 2-fold sample split. The findings further underscore the advantages of MR-SPLIT.

In summary, MR-SPLIT consistently demonstrates robust type I error control when compared to 2SLS and LIML across various simulation settings. In comparison to CFMR, MR-SPLIT exhibits superior performance, yielding smaller RMSE and higher statistical power. Even under a less favorable condition for MR-SPLIT, the type I error can be controlled when the sample size is reasonably large. The simulation results further corroborate our theoretical finding, consistently showing that MR-SPLIT results in smaller standard errors for causal effect estimation compared to CFMR, which leads to higher statistical power when testing for the causal effect.

2.3.4 Multiple Data Splitting

Intuitively, more data splitting should yield more robust results, which, however, would entail higher computational resource usage. We implemented our methods under different splitting times, different sample size N and different h^2 values, to check if we can find an efficient number of splitting. In our simulations, the true causal effect of the exposure on the outcome is set to equal 0.2

($\beta = 0.2$), and the sample size ranges from 500 to 2000 ($N = 500, 1000, 2000$). We did simulations in Scenario I as described in section 2.3.3. Fig 2.5 demonstrates how the type I error fluctuates with an increasing number of splits. To obtain a smoother estimate of the type I error rate, we repeated the simulation 5,000 times Under a small sample size, the type I error rates get stable as the number of sample splits increases. Though the type I error increases as the sample split times increase under small sample sizes, this increase is considered acceptable, particularly in light of the associated boost in power (see Fig 2.5), which is especially pertinent for smaller sample sizes.

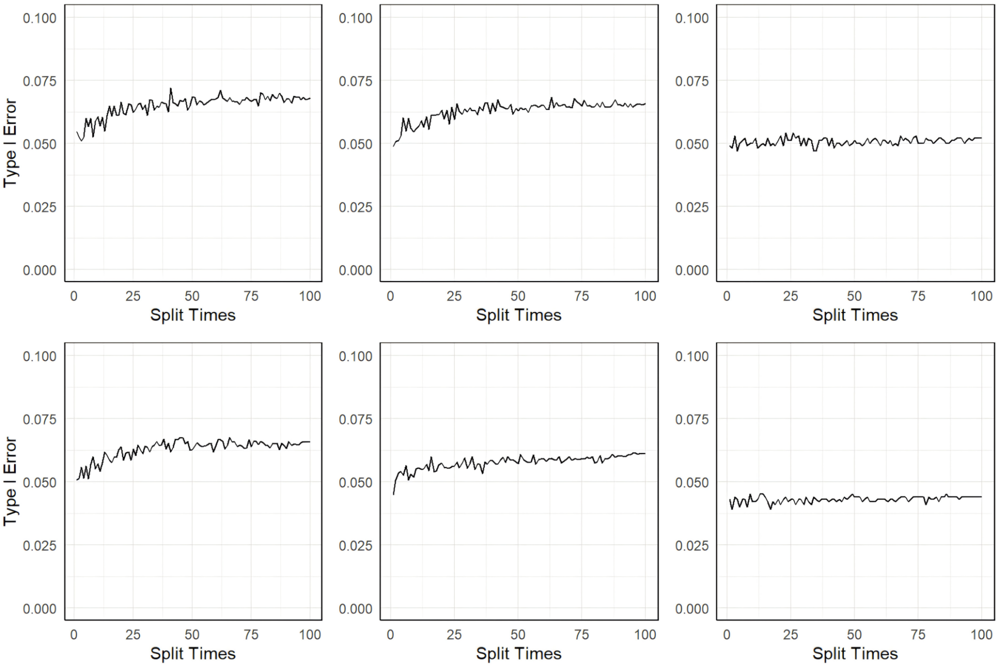


Figure 2.5 Type I error under different sample sizes: N = 500(left), 1000(middle), 2000(right), and under different h^2: 0.15 (top) and 0.2 (bottom).

Fig 2.6 shows the empirical power under different sample sizes and h^2. The type I error and power results when $h^2 = 0.3$ can be found in Figs A.12 and A.13 in Appendix A.2. When the sample size is small ($N = 500$) and the IVs are relatively weak ($h^2 = 0.15$), the power gets stabilized after 50 splits. As the sample size increases, there is a decrease in the need for the number of sample splits to maintain stable power. This indicates that in practical data analysis, it is possible to estimate the exposure heritability based on the selected SNP IVs, and thereafter determine the appropriate number of sample splits. In any case, opting for 50 sample splits represents a highly conservative option.

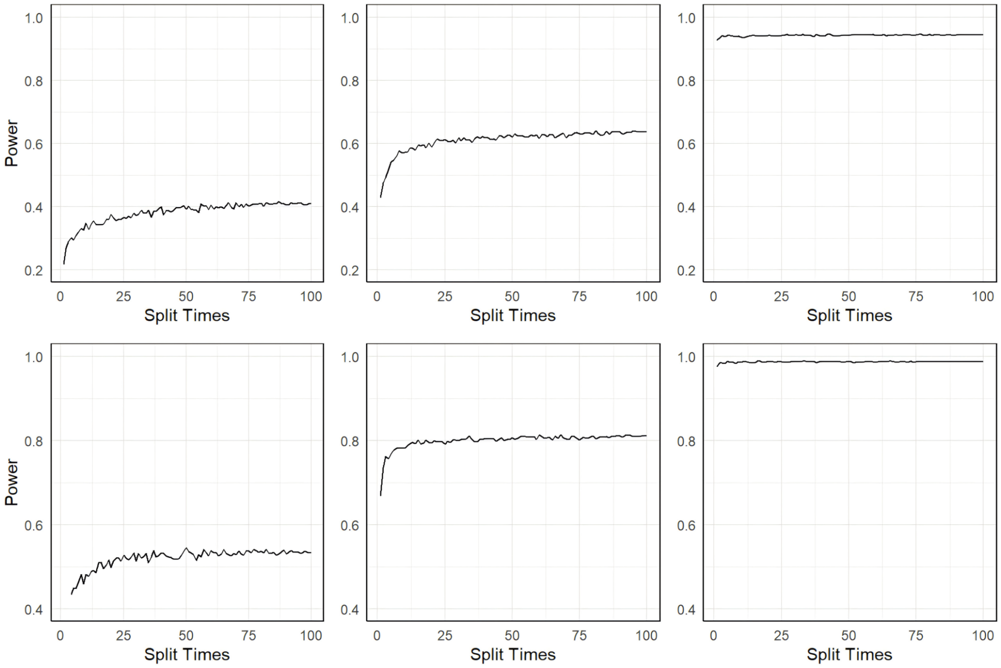


Figure 2.6 Empirical power under different sample sizes: N = 500(left), 1000(middle), 2000(right), and under different h^2: 0.15 (top) and 0.2 (bottom).

2.4 Case Study: eGFR and aTRH, uACR and aTRH

We demonstrated the effectiveness of our method by applying it to the Chronic Renal Insufficiency Cohort (CRIC) dataset, to understand the progression of chronic kidney disease (CKD).

CKD is evaluated utilizing two straightforward tests: a blood test known as the estimated glomerular filtration rate (eGFR) and a urine test, the urine albumin-creatinine ratio (uACR). Both eGFR and uACR measure kidney function, with low eGFR and high uACR values indicating impaired kidney function. In this application, we are interested in evaluating the causal relationship between CKD and apparent Treatment-Resistant Hypertension (aTRH). aTRH is a condition where a patient's blood pressure remains above target levels despite using three different classes of antihypertensive drugs at optimal doses, typically including a diuretic. The definition of aTRH also extends to cases where four or more medications are required to effectively control blood pressure (Judd and Calhoun, 2014). In a recent two-sample MR analysis using summary statistics, Yu et al. (2020) identified the causal effect of higher kidney function (measured by eGFR estimated from serum creatinine) on lower systolic blood pressure. To date, the causal relationship between CKD and aTRH and the causal link between them remains to be established (Chen et al., 2019; Thomas et al., 2016; Kaboré et al., 2017; Kabore et al., 2016). To this end, we utilized eGFR and uACR as the exposure variable and aTRH as the outcome, applying MR-SPLIT for our analysis. For comparative purposes, we also employed CFMR and 2SLS on the same dataset. Given that the outcome variable (aTRH) is binary (0/1) in nature, the LIML method is not suitable in this analysis.

2.4.1 Genetic Data Processing

The original data have 3,541 samples containing 970,342 SNPs. Our initial step involved removing SNPs with missing rate larger than 10%, resulting in 886,384 SNPs. After excluding SNPs with minor allele frequency (MAF) lower than 0.05, 762,664 SNPs were left. The next phase entailed the elimination of SNPs with p-values less than 1e-5 in the Hardy-Weinberg equilibrium test, which narrowed our SNP count down to 693,848. To ensure the robustness of our genetic instruments, we then implemented LD pruning. SNPs were filtered out in close LD by considering pairs of SNPs within a window of 100 kb. If a pair of SNPs has an LD measure (r^2) exceeding 0.64, one SNP from the pair is removed. After completing all these steps, we were left with 467,597 SNPs.

2.4.2 Causal Analysis

2.4.2.1 Causal effect of eGFR on aTRH

In the initial dataset, eGFR values were obtained on multiple occasions. For consistency and relevance, we selected the eGFR measurements corresponding to visit number 3, which also represents the baseline assessment. Following the exclusion of samples with missing values for either eGFR or aTRH, and then combined with the SNP data, our analysis proceeded with a total of $N = 1,353$ samples. A simple logistic regression shows there is a strong association between aTRH and eGFR ($p < 2 \times 10^{-16}$). We would like to evaluate if this association is causal. Fig A.20 shows the boxplots of eGFR in aTRH positive and negative groups.

Next, we proceeded with the MR-SPLIT and used SIS for preliminary screening, reducing the number of SNPs from ultra-high to high. To optimize computational efficiency in the analysis, we first conducted univariate regression of each SNP against the exposure before applying sample split, using the whole data set. A total of 4,580 SNPs ($p < 0.01$) remained for further analysis. The removed SNPs would most likely be screened out by the SIS procedure in subsequent steps even after the sample split if not discarded at this stage. For each of the 50 sample splits, we used the 'screening' function from the R package 'screening' with the SIS option. The number of SNP variables retained post-screening adhered to the default setting, which is half the size of the sample. In this real data analysis, instead of applying the LASSO algorithm to select and estimate SNP effects, we employed a high-dimensional inference procedure, specifically a LASSO-projection method which provides debiased coefficient estimates and hence a valid p-value for each coefficient. This is done by using the 'lasso.proj' function in the R 'hdi' package(Dezeure et al., 2015). As the regular LASSO estimates are biased, this approach can give debiased estimates and further provide p-values for testing each coefficient. To compare the performance of the LASSO-projection with the regular LASSO, We conducted a simulation (detailed in Section A.2.7 in A.2). The results show that the LASSO-projection method slightly outperforms LASSO, exhibiting higher power and better control of the type I error rate and smaller RMSE. After getting the p-values for each SNP, we retained those with a p-value less than or equal to 0.05. This resulted in an average of 98 IVs out of 50 sample splits. We used the partial $F > 30$ as the criterion to declare major IVs. And the weak IVs were then combined into a composite IV. Finally, we used both the composite IV and the major IV(s) to obtain the causal effect estimate and the p-value.

Fig 2.7 shows the p-value distribution and the causal effect estimates out of 50 sample splits. The majority of p-values obtained from MR-SPLIT are below 0.05, and the majority of the estimated causal effects $\hat{\beta}$ is centered around -0.0343 (indicated by the black dashed line). In these 50 sample splits, there was an average of 98.06 IVs incorporated into the model for the causal effect estimate and the majority were classified as weak IVs. Among these, an average of 0.54 IVs were identified as major IVs each time. After aggregating all the results using Cauchy's combination rule, our method provided an estimate of $\hat{\beta} = -0.0343$ (OR= 0.9663), with an aggregated p-value of 5.96×10^{-5}. We also tried lowering the partial F threshold to 20, which yielded slightly more major IVs than the $F > 30$ threshold (see Fig A.22 in Appendix A.2). Among the 50 sample splits, an average of 4.26 IVs were identified as major IVs each time. The results show that the p-value for MR-SPLIT improved slightly (from 5.9×10^{-5} to 2.9×10^{-6}), but the estimates remained nearly the same ($\hat{\beta} = -0.0342$).

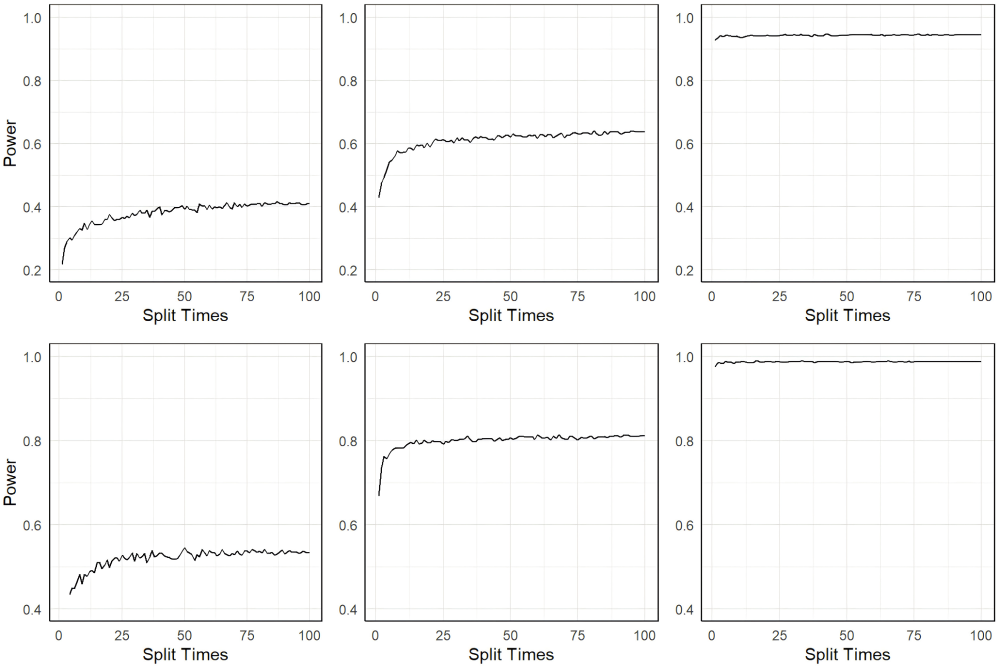


Figure 2.7 Histogram of p-values and causal effect estimates from 50 sample splits when eGFR is treated as the exposure.

We also applied the CFMR method with a 10-fold split. The CFMR method yielded an average estimate of $\hat{\beta} = -0.0378$ (OR=0.9629), with a p-value of $< 1 \times 10^{-5}$. For reference, simply conducting the 2SLS method yields an estimate of $\hat{\beta} = -0.0407$ (OR= 0.9601), with a p-value of $< 1 \times 10^{-7}$. The three methods established a consistent causal relationship between eGFR and aTRH.

2.4.2.2 Causal effect of uACR on aTRH

Following a similar procedure, we excluded samples with incomplete data for either uACR or aTRH. After merging the remaining data with the SNP data, the dataset was reduced to 1,324 samples. The distribution of uACR is very skewed (to the right) (See Fig A.23 in Appendix A.2). We opted to do a logarithmic transformation of uACR, denoted as log(uACR). A simple logistic regression shows there is a strong association between log(uACR) and aTRH ($p = 4.6 \times 10^{-12}$). Similar procedures as described before were followed for further analysis.

Fig 2.8 shows the p-value distribution as well as the causal effect estimate out of 50 sample splits with MR-SPLIT. In these 50 sample splits, there were on average 74.3 SNPs selected as IVs with the majority as weak ones for the causal effect estimate. Among them, an average 0.22 IVs were identified as major IV each time. After aggregating all the results, the final causal estimate was $\hat{\beta} = 0.1675$ (OR= 1.186), with a p-value of 1.9×10^{-3}. For comparison, CFMR provided an estimate of $\hat{\beta} = 0.1584$ (OR= 1.1716), with a p-value of 5.2×10^{-5}. The two methods yielded statistically significant results and presented comparable estimates. While applying 2SLS on the same dataset, we also observed significant results (p-value=1.3×10^{-5}), albeit with a different causal effect estimate of $\hat{\beta} = 0.0363$.

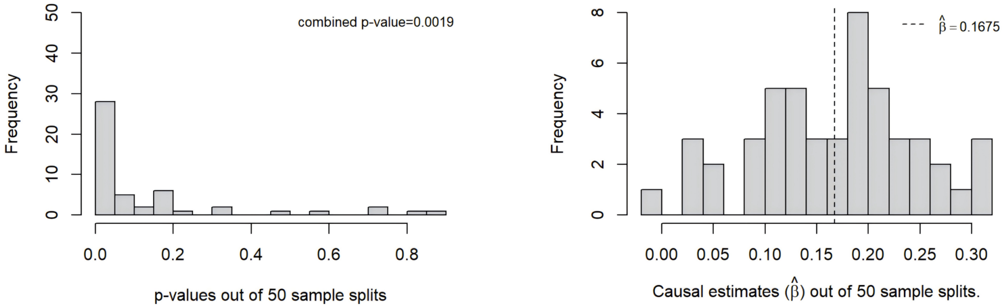


Figure 2.8 Histogram of p-values and causal effect estimates from 50 sample splits when log(uACR) is treated as the exposure.

Integrating the results from the two analyses that utilized eGFR and uACR separately as exposures, we infer that there exists a causal relationship between CKD function and aTRH. Specifically, a lower eGFR and a higher uACR tend to contribute to an increased risk of aTRH. However, we recognize the limited sample size of this study, which necessitates cautious interpretation of the causal relationship identified. To assess the possibility of a reverse causal effect, we require a method capable of accommodating a binary exposure variable, such as aTRH in this context. This will be explored in our future studies.

2.5 Discussion

MR analysis has been an instrumental means in epidemiology studies, enabling the assessment and revelation of causal connections between exposures or interventions and particular outcomes, leveraging genetic variants as IVs to mitigate confounding factors. In this study, we introduced an innovative adaptive sample splitting method known as MR-SPLIT, designed to address the issue of IV selection bias and weak instruments in the context of one-sample MR analysis using individual-level data. By a random sample split, we use half sample to select IVs and another independent half to estimate the causal effect, hence avoiding the winner's curse problem by using the same data for IV selection and causal effect estimation. Additionally, we presented a multi-sample splitting strategy to further enhance the robustness of causal estimation and testing. Our approach involves the adaptive identification of major and weak IVs and further aggregate weak IVs to form a composite IV. The final set of IVs comprises the major IV(s) and the composite IV. Such a strategy, as shown in the theoretical evaluation and simulation results, yields a more efficient causal estimate than CFMR, thereby enhancing testing power. In addition, MR-SPLIT shows consistently superior performance in terms of coverage probability. Therefore, MR-SPLIT offers significant improvements over existing methods by effectively handling weak instruments in one-sample MR analysis and providing robust results with enhanced statistical power.

In comparison to the traditional 2SLS and LIML methods, MR-SPLIT yields less biased results and effectively controls type I error, under different simulation settings. Compared to the CFMR approach, which is designed to tackle weak IV issues, our approach provides estimates with smaller variance and higher statistical power. In the application to the CRIC dataset, both MR-SPLIT and CFMR produce highly comparable results. We established the causal impact of kidney function, as assessed by eGFR and uACR, on aTRH. It is worth noting that both CFMR and MR-SPLIT not only address the issue of weak instrument bias (i.e. finite-sample bias from IV analysis with a given set of IVs), they also solve the problem of "winner's curse" (i.e. bias due to variant selection in the same dataset as the analysis is performed, in particular under a high-dimensional scenario). The two sets of biases are related but are conceptually distinct. By employing sample splitting strategies, both methods tackle the two bias issues and offer a solution to one-sample MR analysis. On the other hand, as shown in our theoretical evaluation as well as the intensive simulation studies, MR-SPLIT demonstrates superior performance compared to CFMR. Within the proposed sample splitting strategy, additional tasks such as nonlinear causal estimation can also be executed using one-sample individual-level data.

In the process of selecting IVs, CFMR recommends employing predictive methodologies, such as LASSO regression, for their efficacy in enhancing prediction accuracy through variance minimization. However, this approach often introduces bias in effect estimates, as it may incorporate SNPs without significant association with the exposure - potentially compromising the relevance assumption for IVs. On the other hand, 2SLS analysis prioritizes the use of predicted exposure values in its secondary causal inference phase, underlining the importance of prediction accuracy for causal estimation. Recent advancements in the realm of high-dimensional statistical inference offer a promising solution by enabling the evaluation of estimation uncertainty for LASSO-derived estimates (Dezeure et al., 2015). This is achieved through a de-biasing step that facilitates the calculation of p-values, thereby presenting an innovative approach for SNP IV selection within the context of high-dimensional SNP-exposure regressions. This technique allows for the derivation of p-values for individual SNPs, enabling the validation of IV suitability through a p-value based method. Unlike traditional practices that determine p-values by fitting each SNP individually in marginal regressions, this approach fits all SNPs (after the SIS step) in a multiple regression model. This yields partial SNP effect estimates and hence partial p-values, offering a nuanced perspective compared to conventional methods. By adopting a p-value threshold criterion (e.g., $p < 0.05$), the selected SNPs meet the relevance assumption, providing a more robust framework for IV selection. In our analysis, we observed that the LASSO variable selection technique typically identifies a greater number of IVs compared to the debiased LASSO method. If computational resources are not a limiting factor, we recommend the implementation of the debiased LASSO approach in practical applications.

While MR-SPLIT offers notable advantages, there is still considerable potential for further enhancement and refinement. In this work, we applied the partial F statistics for identifying major IVs, which does not rule out the application of other measures such as those studied by Stock and Yogo (2002). Any statistical measure capable of ranking the effect sizes of the selected IVs could be considered for enhancing the robustness and effectiveness of our approach. It is essential to devise robust methods for discerning between major and weak IVs. This represents a promising direction for future research. It is worth mentioning that we do not specify the ratio of major IVs to weak IVs; their quantities are entirely determined by the data itself, that is, based on the strength calculated from the IVs. On the other hand, as revealed by the simulation studies, the declaration of major IVs may vary under different F thresholds and under different sample sizes and SNP heritability levels. In real applications, the $F > 30$ threshold can be relaxed under a small sample size and low heritability level. The genomewide SNP heritability can be estimated with software such as GCTA (Yang et al., 2011).

In addition, we employed a straightforward weighted combination approach to aggregate the information from all weak IVs into a single composite IV. Other advanced machine learning techniques could also be borrowed by minimizing information loss which could potentially yield improved results.

An additional constraint of MR-SPLIT is that we did not take the pleiotropic effects into consideration, but there are several test statistics available to identify its presence (Greco M et al., 2015; Sargan, 1958). Studies also show that incorporating the invalid IVs with uncorrelated and correlated horizontal pleiotropic effects can potentially increase power and decrease bias (Yuan et al., 2022; Qi and Chatterjee, 2019; Burgess et al., 2020). We will investigate this in Chapter 3.

The concept of sample splitting and cross-fitting instruments introduced in this study has potential applications beyond the scope of traditional one-sample MR analyses using individual-level data. For example, this framework can be adapted for use in multiple exposure MR analyses, where it would involve adapting the existing approach to handle multiple sets of selected IVs simultaneously. For another example, the proposed framework enables the investigation of potential non-linear causal relationships through a control function approach while effectively addressing the two bias issues previously mentioned. Accomplishing this task is not feasible with summary statistics, highlighting the framework's capability to provide more nuanced insights into causal mechanisms that cannot be captured by summary-level data. In essence, the expansion of our methodology to encompass various types of MR analyses could facilitate innovative research into causal relationships, opening new avenues for investigation.

CHAPTER 3

MR-SPLIT+ — ROBUST CAUSAL INFERENCE WITH MANY WEAK AND INVALID INSTRUMENTS

3.1 Introduction

Depending on the type of data available, MR analysis methods can broadly be categorized into two classes: those based on summary statistics and those utilizing individual-level data. The former have been widely adopted in two-sample MR analysis (Bowden et al., 2015, 2016; Verbanck et al., 2018), primarily because they avoid privacy concerns and allow for easier access to data. However, despite the growing availability of summary statistics, their use presents several inherent limitations. Flexible adjustment for covariates is generally not feasible with summary level data, potentially compromising the precision of causal estimates. Additionally, the lack of individual level data necessitates reliance on external reference panels for LD pruning, which may introduce bias. More complex modeling frameworks, such as nonlinear MR analysis, also require individual level data and are impractical to implement using summary statistics alone. With the increasing availability of large scale individual level datasets, such as the UK Biobank data, conducting reliable MR analyses at the individual level has become increasingly feasible (Millard et al., 2019; Sproviero et al., 2021; Cheng et al., 2024). These datasets provide rich and detailed genetic and phenotypic information, offering unparalleled opportunities for improving the validity and robustness of causal inference.

MR-SPLIT is a method proposed by Shi et al. (Shi et al., 2024), to solve the weak IV issue and also the selection bias in one-sample MR studies. This method provides nearly unbiased estimates when there are many weak IVs. It also ensures the preservation of statistical power while effectively controlling the type I error rate. However, it did not address the invalid IV issue, which is also known as the horizontal pleiotropy in MR analysis. Thus, it is imperative to further improve this method to effectively address the issue of invalid IVs. This advancement would enhance its reliability in practical applications, allowing researchers to apply the method with confidence and reduced concern over assumption violations.

Built upon the MR-SPLIT framework with multiple splitting to address IV selection bias and weak IV bias in one-sample MR analysis, we propose MR-SPLIT+, an enhanced version of MR-SPLIT. It addresses the invalid IV issue by utilizing a mixed integer optimization algorithm introduced by Bertsimas et al. (Bertsimas et al., 2016), and further combines it with the modified Cragg-Donald test (Kolesár, 2018) for testing of overidentifying restrictions. Compared to prior methods (e.g., sisVIVE, CIIV, or WIT), our method demonstrates greater accuracy in identifying valid IVs and substantially reduces estimation bias under the relaxed plurality rule. By incorporating multiple splitting, which enhances estimate robustness and improves reliability, MR-SPLIT+ achieves performance as good as that of the oracle method.

The structure of this chapter is organized as follows. Section 3.2 reviews the UK Biobank, a large scale repository for individual level genetic data, calling for the need to further explore the rich data source for causal inference. We then highlight the importance of using positive and negative controls in MR analysis built upon the large sample size of UKB data. Section 3.3 introduces the model framework. We first briefly review MR-SPLIT and then present MR-SPLIT+ under the two-stage least squares (TSLS or 2SLS) framework. We then discuss how multiple splitting improves estimation accuracy and summarize the methodological framework. Section 3.4 presents simulation results based on a primary scenario that assumes no noise, thus omitting the IV selection step and directly identifying invalid IVs. A more complex setting involving numerous noisy IVs, which reflects real world scenarios where IV selection may introduce selection bias, is provided in Section 3.4.3 in the Appendix for reference. In Section 3.5, we apply MR-SPLIT+ to the UK Biobank dataset and demonstrate its utility through the positive and negative control. We also conduct a mediation analysis to further explore factors mediating the causal pathway. Section 3.6 concludes with a discussion of the method's strengths, limitations, and future directions.

3.2 Motivation and Scientific Questions (UK Biobank)

3.2.1 UK Biobank Dataset Enables Robust Causal Inference

The UK Biobank is a large scale, population based prospective cohort comprising more than 500,000 individuals (Sudlow et al., 2015). It provides extensive genotype and phenotype data, making it an invaluable resource for MR analyses. Participants were genotyped using high density arrays covering over 800,000 markers, including genome-wide SNPs and exome variants, enabling robust instrument selection for MR studies.

In addition to genetic data, UK Biobank offers a wide range of deeply phenotyped traits derived from questionnaires, physical measurements, biochemical assays, and linked electronic health records. The longitudinal follow-up through national health registries, including hospital episodes, cancer diagnoses, and mortality data, facilitates outcome ascertainment across a broad disease spectrum. The large sample size, comprehensive phenotyping, and availability of individual-level data make UK Biobank particularly well suited for one-sample MR frameworks, allowing for refined exposure-outcome modeling, control of pleiotropy, and implementation of sensitivity analyses. These strengths further motivate methodological development focused on MR analysis with individual-level data.

3.2.2 Scientific Questions and Motivation for Method Development

With the large sample size in UKB data, robust evaluation of methodology development becomes feasible. This includes using positive and negative controls to assess the power and robustness of causal inference with one-sample MR analysis. For this purpose, we examined two exposure–outcome pairs in the UKB data. One pair, body mass index (BMI) and diastolic blood pressure (DBP), has been widely supported by previous findings and serves as a positive control(Yusni et al., 2024; He et al., 2000; Linderman et al., 2018). The other, birth weight (BW) and BMI, is considered a negative control, as it is biologically implausible for BMI measured after age 40 to have any causal effect on birth weight.

3.2.2.1 Using positive control to assess the power of different methods

We leveraged the well-established causal relationship between BMI and DBP to evaluate the performance of our method and its counterparts. Rather than aiming to re-establish causality, we used this known association as a benchmark to assess the consistency of causal effect estimates produced by various methods. Building on the availability of rich individual-level data from UKB data, we conducted analyses in two groups: the overall cohort and a younger subset of participants. The large sample size provided sufficient statistical power to perform subgroup comparisons and to examine the stability of method performance across different populations. A robust MR method is expected to yield consistent results between the two groups, thereby reflecting the underlying causal relationship.

The detailed analysis procedure can be found in Section 3.5.1. We applied five methods in total, including ordinary least squares (OLS), naive TSLS, sisVIVE, CIIV, and WIT. The latter three represent recent methodological developments specifically designed to address the presence of invalid IVs. Results are presented in Table 3.2 and Table 3.3 . Although the latter three methods all exhibited strong statistical significance in both groups, their conclusions appear less convincing upon closer examination. For example, while other methods detected the presence of invalid IVs in the 'all group' results, sisVIVE failed to identify any invalid instruments. Furthermore, sisVIVE only provides point estimates without accompanying statistical tests, which greatly limits its utility in practical applications. As for WIT, it exhibited a notable inconsistency: it identified no invalid IVs among 101 candidates in the 'young group', yet detected 51 invalid IVs out of 99 candidates in the 'all group'. Such contradictory findings undermine confidence in the conclusions drawn from WIT. In the case of CIIV, although its results appear relatively consistent, the method relies on the assumption that instruments are sufficiently strong to ensure valid estimation. In our subsequent simulations covering a wider range of scenarios, the performance of CIIV was also found to be unsatisfactory.

3.2.2.2 Using negative control to assess the robustness of different methods

Building on the availability of individual-level data from the UK Biobank, we further designed a negative control analysis to complement the positive control described earlier. Based on established biological knowledge, an individual's BMI measured at the age of 40 or older cannot plausibly influence their own birth weight, implying the absence of a causal effect from BMI to BW. In this analysis, we treated BMI as the exposure and BW as the outcome. Given the lack of a biologically plausible causal pathway, we did not expect to observe a significant causal effect between the two variables. The negative control setting allows us to further evaluate the robustness of different MR methods. In most real-world applications, the true causal relationship between an exposure and an outcome is unknown, making it difficult to assess the validity of MR estimates. However, in this case, the biological implausibility of the exposure-outcome relationship provides a rare opportunity to benchmark method performance using real data with large sample sizes.

The detailed analysis procedure can be found in Section 3.5.2. Results are presented in Table 3.5. We could find that although WIT is specifically designed to handle the presence of invalid IVs among many weak instruments, it nonetheless produced counterintuitive results, suggesting a false causal effect of BMI on birth weight.

Therefore, based on the findings from the two examples above, we recognize an urgent and essential need to develop a method capable of providing reliable causal estimates using individual-level data. Ideally, this method should be sufficiently robust to accommodate a wide range of practical scenarios, easily interpretable, and preferably built upon the widely used 2SLS framework. Motivated by these goals, we extend the original MR-SPLIT method and propose MR-SPLIT+.

3.3 Methods

Let $Y \in \mathbb{R}^{N\times 1}$ be the outcome variable of interest, and $X \in \mathbb{R}^{N\times 1}$ the exposure variable, where N denotes the sample size. Both Y and X are assumed to be continuous variables. We define $G \in \mathbb{R}^{N\times p}$ as the genetic instruments (i.e., SNPs), where p is the number of SNPs. We futher denote the unknown confounder as U, which could have effects on both X and Y but unobserved. The model can be represented as:

$$\begin{aligned} U &= G\eta_1 + \varepsilon_1, \\ X &= G\eta_2 + U\eta_3 + \varepsilon_2 \\ &= G(\eta_2 + \eta_3\eta_1) + (\varepsilon_2 + \varepsilon_1\eta_3), \\ Y &= X\beta + G\eta_4 + U\eta_5 + \varepsilon_3 \\ &= X\beta + G(\eta_4 + \eta_5\eta_1) + (\varepsilon_3 + \varepsilon_1\eta_5) \end{aligned} \tag{3.1}$$

We call IVs are invalid if $\eta_4 + \eta_5\eta_1 \neq 0$. This setting introduces challenges for causal inference, as standard IV methods assume all instruments affect the outcome solely through the exposure (i.e., $\eta_4 + \eta_5\eta_1 = 0$). Let $\gamma = \eta_2 + \eta_3\eta_1$, $\alpha = \eta_4 + \eta_5\eta_1$, $\varepsilon_x = \varepsilon_2 + \varepsilon_1\eta_3$ and $\varepsilon_y = \varepsilon_3 + \varepsilon_1\eta_5$, the model is simplified as follows,

$$Y = X\beta + G\alpha + \varepsilon_y, \tag{3.2}$$

$$X = G\gamma + \varepsilon_x, \tag{3.3}$$

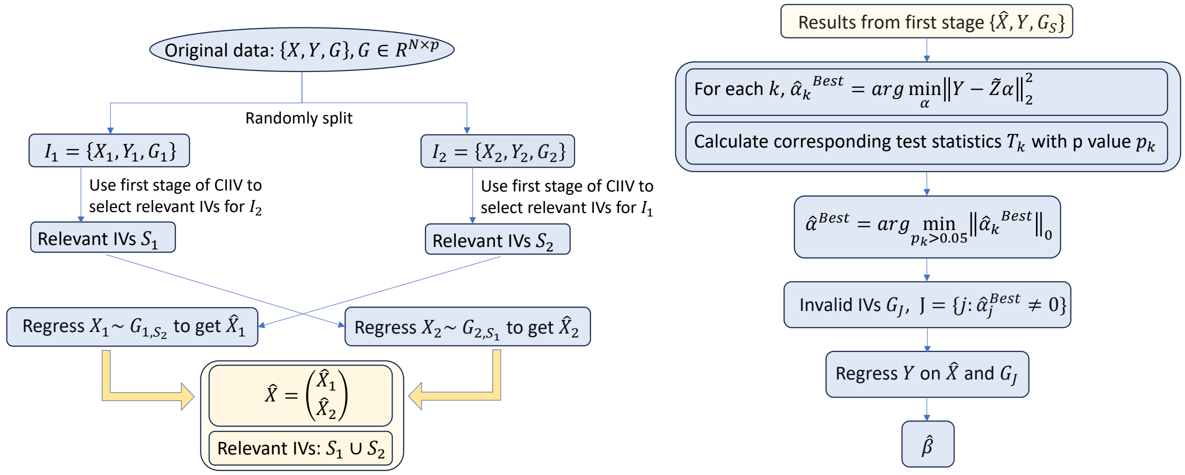
where ε_x and ε_y are error terms assumed to follow normal distributions with mean 0 and $\text{cor}(\varepsilon_x, \varepsilon_y) \neq 0$ due to the influence of unknown confounders. Invalid IVs are indicated by $\alpha \neq 0$. Our objective is to develop a robust framework for estimating the causal effect β, while simultaneously identifying and accounting for invalid instruments to mitigate bias and improve inference accuracy.

3.3.1 MR-SPLIT Recap

Before introducing MR-SPLIT+, we first briefly introduce the framework of MR-SPLIT (Shi et al., 2024), which was designed to solve the weak IV and selection bias issues in one-sample MR studies. Given the observed data $\{X, Y, G\}$, a screening method is first applied, such as SIS (Fan and Lv, 2008), to reduce the number of SNPs from an ultra-high dimension to a more manageable level as the number of SNPs is usually in the magnitude of 10^5 or higher. Next, the data sample is randomly split into two parts. One part is used to select IVs using a shrinkage method such as LASSO. These IVs are then categorized into major and weak ones, based on the partial F-statistics, followed by combining only the weak IVs into a composite one using a weighted approach and obtaining the cross-fitted exposure. The same procedure is applied to the other half of the data. The two sets of cross-fitted exposures are then combined together to fit the second stage IV regression model with the entire sample and to estimate the causal effect. Finally, multiple splits are applied, and the final estimate is defined as the mean of the estimates obtained from multiple splits. The final p-value is aggregated through the Cauchy combination test.

3.3.2 The first stage of MR-SPLIT+

MR-SPLIT has been shown to perform well when there are no invalid IVs. In this work, we propose a rigorous approach to address the invalid IV issues to improve the MR-SPLIT framework. The first stage of MR-SPLIT+ is illustrated in Figure 3.1a. Similar to the MR-SPLIT approach and given the observed dataset $\{X, Y, G\}$, we employ screening methods, such as SIS, to reduce the number of SNP IVs to a manageable size, typically a few hundred.



Original data: $\{X, Y, G\}, G \in R^{N\times p}$
Randomly split
$I_1 = \{X_1, Y_1, G_1\}$
$I_2 = \{X_2, Y_2, G_2\}$
Use first stage of CIIV to select relevant IVs for I_2
Use first stage of CIIV to select relevant IVs for I_1
Relevant IVs S_1
Relevant IVs S_2
Regress $X_1 \sim G_{1,S_2}$ to get $\hat{X}_1$
Regress $X_2 \sim G_{2,S_1}$ to get $\hat{X}_2$
$\hat{X} = \begin{pmatrix}\hat{X}_1\\ \hat{X}_2\end{pmatrix}$
Relevant IVs: $S_1 \cup S_2$
Results from first stage $\{\hat{X}, Y, G_S\}$
For each k, $\hat{\alpha}_k^{Best} = arg \min_{\alpha} \|Y - \tilde{Z}\alpha\|_2^2$
Calculate corresponding test statistics T_k with p value p_k
$\hat{\alpha}^{Best} = arg \min_{p_k > 0.05} \|\hat{\alpha}_k{}^{Best}\|_0$
Invalid IVs G_J, $J = \{j: \hat{\alpha}_j^{Best} \neq 0\}$
Regress Y on $\hat{X}$ and G_J
$\hat{\beta}$

(a) First stage of MR-SPLIT+.

(b) Second stage of MR-SPLIT+.

Figure 3.1 The two stage MR-SPLIT+ framework.

Next, we split the sample evenly into two subsets and conduct IV selection independently in each subsample. Regardless of the method researchers choose for IV selection, such as marginal p-values, LASSO, or adaptive LASSO, we strongly recommend applying the first stage of the CIIV method to further refine the selected IVs (see details in section 3.3.3.4). Across multiple simulation studies, this additional step has proven highly effective in mitigating noise while retaining valid IVs. Since excessive noise often introduces bias in the identification of invalid IVs, incorporating this refinement can significantly improve estimation accuracy.

After completing IV selection in each subset, the key difference between MR-SPLIT+ from MR-SPLIT lies in how the selected IVs are treated. Instead of treating major and weak IVs separately, all IVs are combined into one composite IV. Extensive simulation studies have demonstrated that dealing with one composite IV leads to better type I error control and more accurate coverage probabilities. This approach makes practical sense as SNPs usually have weak effects in GWAS studies, especially under small sample sizes. In the presence of strong IV effect, major IVs can be retained and dealt separately from the weak ones following the MR-SPLIT procedure. We have incorporated this option in our code, allowing users to achieve this by varying the threshold of partial F-statistics. After the above steps, as in the MR-SPLIT approach, we obtain cross-fitted exposures in the two subsamples denoted as $\hat{X}_1$ and $\hat{X}_2$, then combine them to get $\hat{X}$ for the entire sample.

3.3.3 The second stage of MR-SPLIT+

Assuming that some IVs may be invalid, meaning that the IVs may have a direct effect on Y, i.e. $\alpha \neq 0$. In the second stage, our primary goal is to accurately identify the invalid IVs selected in the first stage and include them as covariates in the model to obtain an unbiased causal estimate. Suppose we obtain the selected IV set $G_S = G_{S_1} \cup G_{S_2}$ in the first stage.

3.3.3.1 Identifiability of parameters

Intuitively, one might attempt to use a shrinkage method to solve the following function to identify the invalid IVs:

$$\hat{\alpha} = \arg\min_{\alpha} \|Y - X\beta - G\alpha\|_2^2, \tag{3.4}$$

and if $\hat{\alpha}_j = 0$, G_j is considered a valid IV; otherwise, if $\hat{\alpha}_j \neq 0$, G_j is deemed an invalid IV. However, we noticed that for any constant c, we can rewrite Eq. (4.1) as

$$\begin{aligned} Y &= X\beta + G\alpha + \varepsilon_y \\ &= X(\beta + c) + G\alpha + \varepsilon_y - c(G\gamma + \varepsilon_x) \\ &= X(\beta + c) + G(\alpha - \gamma c) + (\varepsilon_y - c\varepsilon_x) \end{aligned} \tag{3.5}$$

Hence, for every value $c = \alpha_j/\gamma_j \neq 0, j = 1, \cdots, p$, we could use the estimated parameter $\{\beta + c, \alpha - \gamma c\}$ to get the same Y, and each corresponds to a specific set of invalid IVs, characterized by distinct values of α_j. To ensure parameter identifiability, we impose the Plurality Rule proposed by Guo et al. (2018), stated as follows:

Assumption 1 (Plurality Rule)**.**

$$\left|\{\alpha_j : \alpha_j = 0\}\right| > \max_{c \neq 0} \left|\left\{\alpha_j : \frac{\alpha_j}{\gamma_j} = c\right\}\right|. \tag{3.6}$$

This condition ensures that the majority of the instruments are valid, meaning that among all IV subsets classified by different values of $c = \frac{\alpha_j}{\gamma_j}$, the subset of valid IVs is the largest. In fact, this condition can be further relaxed, as we will discuss in detail in Section 3.3.3.2.

Now we can successfully identify the unique set of valid IVs by solving

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|Y - X\beta - G\alpha\|_2^2 < \delta, \tag{3.7}$$

where the ℓ_0 norm of the vector α counts the number of nonzeros in α, and δ is a sufficiently small prespecified value. In other words, among all possible solution combinations, the true values of the parameters are the ones that make α the sparsest.

Following the work by Lin et al. (2024), we could reformulate Eq. 3.7 to obtain the solutions as:

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|Y - \widetilde{G}\alpha\|_2^2 < \delta, \tag{3.8}$$

where $\widetilde{G} = M_{\hat{X}}G = (I - \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}')G \in \mathbb{R}^{N\times\tilde{p}}$. In our method, $\hat{X}$ represents the estimated exposure obtained from the first stage, while G denotes G_S, the union of IVs selected during the first stage.

3.3.3.2 Best subset selection to identify invalid IVs

To solve (4.3), we first identify several candidate solution sets and then select the one that yields the sparsest α among these candidates. Given $k = \|\alpha\|_0$, our goal is to solve the following optimization problem,

$$\min_{\alpha} \|Y - \widetilde{G}\alpha\|_2^2, \tag{3.9}$$

and this is the so-called best subset selection problem (Miller, 2002). The cardinality constraint $||\alpha||_0 = k$ makes problem NP-hard. To avoid this intractable problem, previous methods opted to use surrogate penalty functions for solutions. This is also why the WIT method (Lin et al., 2024) chose to employ the MCP penalty as an alternative. However, this approach of solving the problem using an alternative method instead of directly addressing it often entails potential issues, which can also be observed in the results of our subsequent simulations.

Denote $\hat{\alpha}^{or}$ be the oracle estimator if we know the true invalid IV set in prior, then we have

$$\hat{\alpha}^{or} = (\widetilde{G}_{A_0}^T \widetilde{G}_{A_0})^{-1} \widetilde{G}_{A_0} Y, \tag{3.10}$$

where $A_0 = \{j : \alpha_j \neq 0\}$ and $\widetilde{G}_{A_0}$ is a submatrix of $\widetilde{G}$. This aligns precisely with the form of the OLS estimator. Before we establish the selection consistency, we need an important assumption stated below.

Assumption 2 (Necessary Condition for Selection Consistency)**.** *There exists a constant $d_1 > 0$ such that:*

$$C_{\min}(\alpha, \widetilde{G}) \geq \frac{d_1 \sigma^2 \log p}{n},$$

where $C_{\min}(\alpha, \widetilde{G}) \equiv \min_{\{\alpha_A : A \neq A_0, |A| \leq |A_0|\}} \frac{1}{n \max(|A_0 \setminus A|, 1)} \|\widetilde{G}_{A_0}\alpha_{A_0} - \widetilde{G}_A \alpha_A\|^2$. A_0 is the true invalid IVs set. $|A_0 \setminus A|$ denotes the number of IVs mistakenly omitted from the true invalid IVs set.

The following Theorem 2 guarantees the selection consistency of our method.

Theorem 2. *Suppose $\hat{\alpha}$ is the global minimizer of the following optimization problem:*

$$\min_{\alpha} \|Y - \widetilde{G}\alpha\|_2^2 \quad s.t. \quad \|\alpha\|_0 \leq k,$$

where the residual term $\widetilde{\varepsilon} = Y - \widetilde{G}\alpha$ follows a normal distribution $N(0, \sigma^2 I)$. Denote $A_0 = \{j : \alpha_j \neq 0\}$ and $\hat{A} = \{j : \hat{\alpha}_j \neq 0\}$. If $k = |A_0|$ and Assumption 2 holds, then

$$P(\hat{A} \neq A_0, \hat{\alpha} \neq \hat{\alpha}^{or}) \to 0 \quad as\ n, p \to \infty.$$

Shen et al. (2012, 2013) demonstrated that under Assumption 2, the constrained ℓ_0-method guarantees selection consistency and oracle parameter estimation, where the estimator is shown to consistently select the correct variables and converge to the oracle OLS estimator under Assumption 2. This assumption ensures a minimal degree of separation necessary for correctly identifying invalid IVs, and serves as a fundamental condition under the L_2 metric for any variable selection method, including LASSO, SCAD, or MCP. The proof of Theorem 2 directly follows the work of Shen et al. (2012, 2013).

As shown in Theorem 2, when the number of valid IVs $p - k$ is correctly specified, even when there exists a group of invalid IVs whose number equals that of the valid IVs, our method can always achieve selection consistency under certain assumptions. So we only require a relaxed version of the Plurality Rule, stated as follows:

Assumption 3 (Relaxed Plurality Rule)**.**

$$\left|\{\alpha_j : \alpha_j = 0\}\right| \geq \max_{c \neq 0} \left|\left\{\alpha_j : \frac{\alpha_j}{\gamma_j} = c\right\}\right|. \tag{3.11}$$

Compared to the original Plurality Rule, which requires that the number of valid IVs strictly exceeds that of any group of invalid IVs (classified by distinct values of $c = \frac{\alpha_j}{\gamma_j}$), the relaxed version allows for ties in group sizes. That is, the valid IV group is permitted to have the same cardinality as one or more invalid IV groups.

In our work, we apply the mixed integer optimization (MIO) approach(Bertsimas et al., 2016) to obtain the global minimizer of (3.9) subject to the cardinality constraint. There is an R package available that can implement this approach, provided at `https://github.com/ryantibs/best-subset`. The general MIO problem can be formulated as:

$$\begin{aligned} \min_{\alpha} \quad & \alpha^T Q \alpha + \alpha^T a \\ \text{s.t.} \quad & A\alpha \leq b, \\ & \alpha_i \in \{0, 1\}, \quad i \in \mathcal{I}, \\ & \alpha_j \geq 0, \quad j \notin \mathcal{I}, \end{aligned} \tag{3.12}$$

where $a \in \mathbb{R}^m$, $A \in \mathbb{R}^{k \times m}$, $b \in \mathbb{R}^k$, $Q \in \mathbb{R}^{m \times m}$ and Q is positive semidefinite. $\alpha \in \mathbb{R}^m$ contains both discrete ($\alpha_i, i \in \mathcal{I}$) and continuous ($\alpha_i, i \notin \mathcal{I}$) variables, with $\mathcal{I} \subset \{1, \ldots, m\}$.

Following (3.12), we can reformulate the minimization problem in (3.9) as:

$$\begin{aligned} \min_{\alpha, z} \quad & \alpha^\top (\widetilde{G}^\top \widetilde{G}) \alpha - 2\alpha^\top \widetilde{G}^\top Y + \|Y\|_2^2 \\ \text{s.t.} \quad & (1 - z_i)\alpha_i = 0, \\ & z_i \in \{0, 1\}, \\ & \sum_{i=1}^{\tilde{p}} z_i \leq k, \\ & -\mathcal{M}_U \leq \alpha_i \leq \mathcal{M}_U, \\ & \|\alpha\|_1 \leq \mathcal{M}_l, \end{aligned} \tag{3.13}$$

where z_i indicates whether $\alpha_i \neq 0$, with $\sum_{i=1}^{\tilde{p}} z_i$ representing the number of nonzero elements in α. If $z_i = 0$, then $\tilde{G}_i$ is excluded from the model, implying that G_i is a valid IV. Consequently, $\sum_{i=1}^{\tilde{p}} z_i$

denotes the number of invalid IVs among the selected IV sets. $\mathcal{M}_U$ is a constant such that if $\hat{\alpha}$ is a minimizer of (3.9), then $\mathcal{M}_U \geq \|\alpha\|_\infty = \max|\alpha_i|$. The presence of $\mathcal{M}_U$ and $\mathcal{M}_l$ could improve the performance of MIO. There are additional representations of (3.9) discussed in (Bertsimas et al., 2016), each tailored to different scenarios. The reformulation in (3.13) presented here is particularly useful when $N > \tilde{p}$ and $\tilde{p}$ is on the order of hundreds. Bertsimas et al.(Bertsimas et al., 2016) introduced three methods to estimate $\mathcal{M}_U$ and $\mathcal{M}_l$. Here, we primarily focus on the third method, parameter specifications from advanced warm-starts.

Consider the following optimization problem:

$$\min_{\alpha} g(\alpha) \quad \text{subject to} \quad \|\alpha\|_0 \leq k, \tag{3.14}$$

where $g(\alpha) \geq 0$ is convex and has a Lipschitz continuous gradient, i.e., $\|\nabla g(\alpha) - \nabla g(\tilde{\alpha})\| \leq \ell\|\alpha - \tilde{\alpha}\|$, with ℓ being the Lipschitz constant.

The following Algorithm 3.1 outlines the steps to provide solutions for (3.14).

Algorithm 3.1 Find a stationary point of problem (3.14).

```
Input: $g(\alpha)$, parameter $L > l$, and the convergence tolerance $\epsilon$.
Output: A first-order stationary solution $\alpha^*$.
1: Initialization with $\alpha_1 \in \mathbb{R}^p$ such that $\|\alpha_1\|_0 \leq k$.
2: For $m \geq 1$
       $\alpha_{m+1} = \lambda_m \eta_m + (1 - \lambda_m)\alpha_m,$
   where $\eta_m \in \mathcal{H}_k\left(\alpha_m - \frac{1}{L}\nabla g(\alpha_m)\right)$, with $\lambda_m \in \arg\min_\lambda g(\lambda\eta_m + (1 - \lambda)\alpha_m)$. The operator
   $\mathcal{H}_k(c)$ is defined component-wise as:
       $\mathcal{H}_k(c)_i = \begin{cases} c_i, & \text{if } i \in \{1, \ldots, k\}, \\ 0, & \text{otherwise.} \end{cases}$
   Here, $\{1, \ldots, k\}$ represents the indices of the $k$ largest absolute values of the vector $c$.
3: Repeat step 2 until $g(\alpha_m) - g(\alpha_{m+1}) \leq \epsilon$.
```

Once we obtain the estimated $\hat{\alpha}$ for (3.14), setting $\mathcal{M}_U := \tau\|\hat{\alpha}\|_\infty$, where τ is a multiplier greater than 1 (e.g., $\tau \in \{1.5, 2, 5\}$), provides a suitable estimate for the parameter $\mathcal{M}_U$. Additionally, defining $\mathcal{M}_l = k\mathcal{M}_U$ yields a reasonable upper bound for $\|\alpha\|_1$. These estimation processes can all be implemented using the `bs()` function from the R package `bestsubset`.

3.3.3.3 Test of over-identification

For each $k = \|\alpha\|_0$, we obtain a potential set of invalid IVs by solving the best subset problem in Section 3.3.3.2. The next step is to determine which of these sets are acceptable, i.e., those that make $\|Y - \widetilde{G}\alpha\|_2^2$ sufficiently small. Instead of specifying a sufficiently small threshold δ, we adopt a testing-based approach proposed by Kolesár (2018).

Denote Z as the selected valid IV set and W as the selected invalid IV set. We rewrite Eqs. (4.1) and (4.6) as following:

$$\begin{bmatrix} X & Y \end{bmatrix} = \begin{bmatrix} Z & W \end{bmatrix} \begin{bmatrix} \gamma & \Gamma \\ \Psi_1 & \Psi_2 \end{bmatrix} + \begin{bmatrix} V_1 & V_2 \end{bmatrix}. \tag{3.15}$$

The hypothesis we aim to test is the Proportionality Restriction (PR) assumption introduced by Kolesár (2018), which is stated as follows:

Assumption 4 (Proportionality Restriction)**.** $\Gamma = \gamma\beta$.

Consider the following statistics:

$$S = \frac{1}{N-k-l} Y'(\mathbf{1}_N - ZZ' - W(W'W)^{-1}W)Y,$$

$$T = \frac{1}{N} Y'ZZ'Y,$$

where k and l are the number of valid and invalid IVs respectively, and N is the sample size. We have $E(T - \frac{k}{N}S) = \mathcal{X}$, where $\mathcal{X} = \frac{1}{N}\begin{pmatrix}\Gamma & \gamma\end{pmatrix}'\begin{pmatrix}\Gamma & \gamma\end{pmatrix}$. Under Assumption 4, we have:

$$\mathcal{X} = \mathcal{X}_{22} \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix},$$

and $\mathcal{X}_{22}$ is the bottom-right submatrix of $\mathcal{X}$. Therefore, testing Assumption 4 is equivalent to testing the following hypotheses:

$$\text{H}_0\text{: } \mathcal{X} \text{ is reduced rank,} \quad \text{v.s.} \quad \text{H}_1\text{: } \mathcal{X} \text{ is positive definite.}$$

Let λ_{min} denote the minimum eigenvalue of the matrix $S^{-1}T$. The test statistic is defined as:

$$\hat{J}_{MD} = \begin{cases} 0, & \text{if} \quad \lambda_{min} \leq \frac{k}{N}, \\ \left(\lambda_{min} - \frac{k}{N}\right)^2, & \text{otherwise.} \end{cases}$$

The specific distribution of this statistic can be found in Kolesár (2018). Moreover, the readily available R package `manyIV` can be employed for this purpose.

3.3.3.4 Refine IV selection with CIIV

In practical applications of one-sample MR analysis, erroneous inclusion of SNPs that do not affect the exposure (i.e., noise) as IVs is a common issue, particularly when the sample size is small. Excessive noise in the candidate IV set can significantly reduce the accuracy of identifying invalid IVs. To address this, we adopt the first-stage filtering procedure from the CIIV method (Windmeijer et al., 2021), which can effectively filter out noises. While the CIIV first-stage filtering is intended to exclude uninformative IVs, it is important to note that it does not exclusively retain strong IVs; weak IVs may also pass through. Nevertheless, this approach provides several advantages:

- It minimizes the inclusion of noises.
- By imposing a less stringent threshold, it avoids selecting only the strongest IVs, allowing for a balanced selection of strong and weak IVs. This flexibility provides room for addressing the weak instrument problem using MR-SPLIT+, reducing bias compared to directly applying CIIV.

Given these benefits, the first-stage filtering procedure is a necessary step for selecting relevant IVs in one-sample MR analyses. Specifically, the first stage aims to test $H_0 : \gamma_j = 0,\ j = 1, \ldots, p$, where p is the number of selected SNPs. We reject the H_0 if

$$|t_{\gamma_j}| = \left| \frac{\hat{\gamma}_j}{\sqrt{\text{var}(\hat{\gamma}_j)}} \right| < \omega_N, \tag{3.16}$$

where $\omega_N = \sqrt{2.01 \log\{\max(p, N)\}}$, and $\text{var}(\hat{\gamma}_j)$ can be a robust variance estimator in case of heteroskedasticity. The framework for the second stage of MR-SPLIT+ is summarized in Figure 3.1b.

3.3.4 Multiple sample splitting to enhance stability and robustness

Due to the uncertainty in selecting valid IVs when splitting a sample only once, we implement a multiple splitting strategy in MR-SPLIT+, to ensure robust results. Suppose we split the sample T times, obtaining an estimate of β_t at each time of split. We choose the estimate that is closest to the median of the set $\{\beta_t : t = 1, \ldots, T\}$ as our final causal effect estimate (see Algorithm 3.2 for the detail). The primary reason we select a single result instead of integrating all split results is that the process of screening invalid IVs often produces outliers. To ensure robustness, we opt for the median-type estimate. This also motivates us to avoid using the p-value combination method employed in MR-SPLIT, and instead use the p-value corresponding to the final estimate as the p-value for testing the causal effect.

We evaluated the effectiveness of multiple sample splitting under the simulation settings described in Section 3.4.3 considering the case with noise IVs. The results, presented in the the Appendix, indicate that performing multiple splits significantly improves the precision of the estimates compared to using a single split, particularly in scenarios with limited sample sizes or substantial noise.

Algorithm 3.2 summarizes the analytical procedure of MR-SPLIT+.

3.4 Simulation Study

3.4.1 The impact of sample split on the effect estimate

We first evaluated the effectiveness of multiple sample splitting under the simulation settings described in Section 3.4.3 considering the case with noises. The results indicate that performing multiple splits significantly improves the precision of the estimate compared to a single split, particularly in scenarios with limited sample sizes or substantial noise. Figure 3.2 presents the violin plots of the estimation results under various split times from 1 to 30. Each plot shows the results under 1000 simulation runs. We also conducted simulations with $N = 6000$, and the detailed results can be found in Fig. A.24 in the Appendix. When the sample size is large and the IVs are strong (Case 1 and Case 2), increasing the number of splits has a limited effect on improving estimation accuracy. However, when IVs are weak or the sample size is small (Case 3 and Case 4),

Algorithm 3.2 The analytical procedure of MR-SPLIT+.

```
Input: $\{X, Y, G\}$, $X, Y \in \mathbb{R}^{N\times 1}$, $G \in \mathbb{R}^{N\times p}$, $p \gg N$.
Output: The causal effect $\hat{\beta}$.
1: Employ screening methods, such as SIS, to reduce the number of SNPs to a manageable size, typically a few hundred.
2: For $t = 1, \dots, T$
     • Split the sample evenly into two subsets $\{I_1, I_2\}$.
     • In subset $I_1$, after selecting a candidate IV set, we apply the first stage of CIIV to refine the IVs and estimate their effects. Then, in $I_2$, the selected IVs are combined into a composite IV using the estimated effects as weights. The same process is repeated for another subset.
     • Get estimated $\hat{X}_1$ and $\hat{X}_2$ from the two subset and combine them into $\hat{X}$. Denote $G_S \in \mathbb{R}^{N\times\tilde{p}}$ as the union set of the two selected IV sets.
     • Starting from $k = 0$, set $\alpha_k = \arg\min_\alpha \|Y - \widetilde{G}\alpha\|_2^2$, then perform the overidentification test. If the testing p-value $\leq 0.05$, increment $k$ by 1 ($k = k + 1$) and repeat the process until the testing p-value $p > 0.05$.
     • Regress Y on $\hat{X}$ while including the selected invalid IVs as covariates to get the estimated causal effect $\hat{\beta}_t$.
3: Perform the same sample split procedure $T$ times and choose
     $$\hat{\beta} = \arg \min_{\hat{\beta}_t \in \{\hat{\beta}_t : t=1,\dots,T\}} \left|\hat{\beta}_t - \text{median}(\{\hat{\beta}_t : t = 1, \dots, T\})\right|$$
   as the final causal effect estimate.
```

the estimates become increasingly stable as the number of splits increases, with a gradual reduction in variance and outliers gradually vanishing.

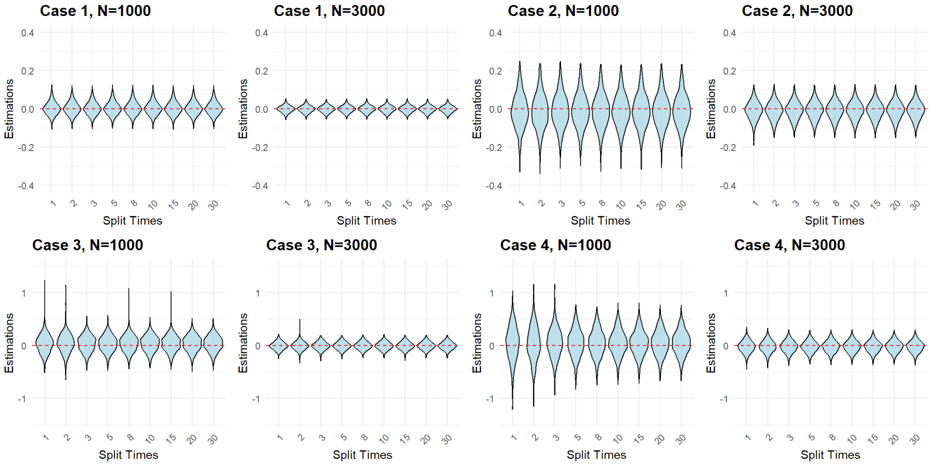


Figure 3.2 Violin plots showing the estimation accuracy as the number of sample split increases under different cases and sample sizes.

Additionally, we examined the changes in coverage probability as the number of sample splits increases (see Figure 3.3). The results indicate that the coverage probability improves as the sample size increases, under different cases. As the number of sample splits increases, the coverage probability shows significant improvement in Case 4 when the sample size is small (e.g., 1000). Under different cases, the coverage probability stabilizes with less than 10 sample splits, indicating the robustness of the method. under weak IVs or small sample sizes, increasing the split times significantly improves the coverage probability, bringing it closer to the nominal 95% level. However, an interesting observation arises in Case 1 with strong IVs: as the sample size increases, the coverage probability approaches the nominal 95% level at N = 3000. However, when N further increases to 6000, the coverage probability instead decreases to around 94%. This phenomenon may be attributed to slight inaccuracies in the variance estimation of our method. As the sample size grows, the estimated variance decreases, leading to a narrower confidence interval. If the variance is slightly underestimated, the confidence interval may become too narrow, resulting in a coverage probability below the nominal level. This suggests a potential area for methodological refinement. Nevertheless, given the small deviation, we still consider this result to be within an acceptable range.

We also presented the results of False Negative Rate (FNR) and False Positive Rate (FPR) for identifying invalid IVs in Fig. 3.4. The FNR represents the proportion of invalid IVs incorrectly identified as valid ones, while the FPR is the proportion of valid IVs incorrectly identified as invalid ones. Similarly, in Case 1 and Case 2, increasing the number of splits did not lead to substantial improvements in performance. However, in Case 3 and Case 4, where the instruments are relatively weak, both FPR and FNR exhibited a decreasing trend as the number of splits increased.

In summary, according to the simulation results, even with extremely weak IVs (Case 4), splitting the sample up to 20-30 times is sufficient to achieve stable results. If the sample size is large, the number of splits can be reduced.

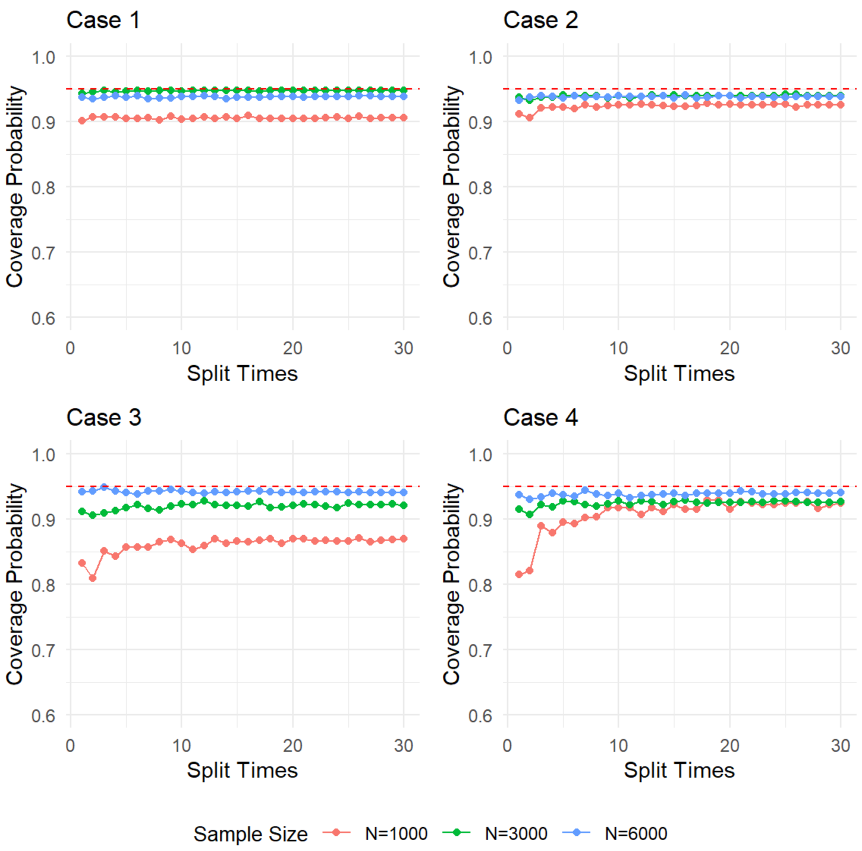


Figure 3.3 Coverage probability under different cases and sample sizes as the number of sample splits increases from 0 to 30.

3.4.2 Simulation without noise

We compared MR-SPLIT+ with recently developed one-sample MR methods, i.e., sisVIVE (Kang et al., 2016), CIIV (Windmeijer et al., 2021), WIT (Lin et al., 2024), as well as the oracle TSLS method which assumes that we know which IVs are valid and invalid in advance.

To ensure a fair comparison, we strictly adhered to nearly all the parameter settings outlined in the WIT work (Lin et al., 2024). The only difference is that we assumed IVs are independent of each other, an assumption that can be easily satisfied in real data through LD pruning. Specifically, to generate data containing N samples, we assumed $G \overset{\text{i.i.d.}}{\sim} N(0, \Sigma^G)$, where $\Sigma^G_{ii} = 0.8, i = 1, ..., p$, and $\Sigma^G_{ij} = 0, i \neq j$. In this section, we assumed that all 21 IVs affect the exposure X, with no noise IVs included. We also considered scenarios that include noise IVs and applied these methods to select IVs as shown in Section 3.4.3.

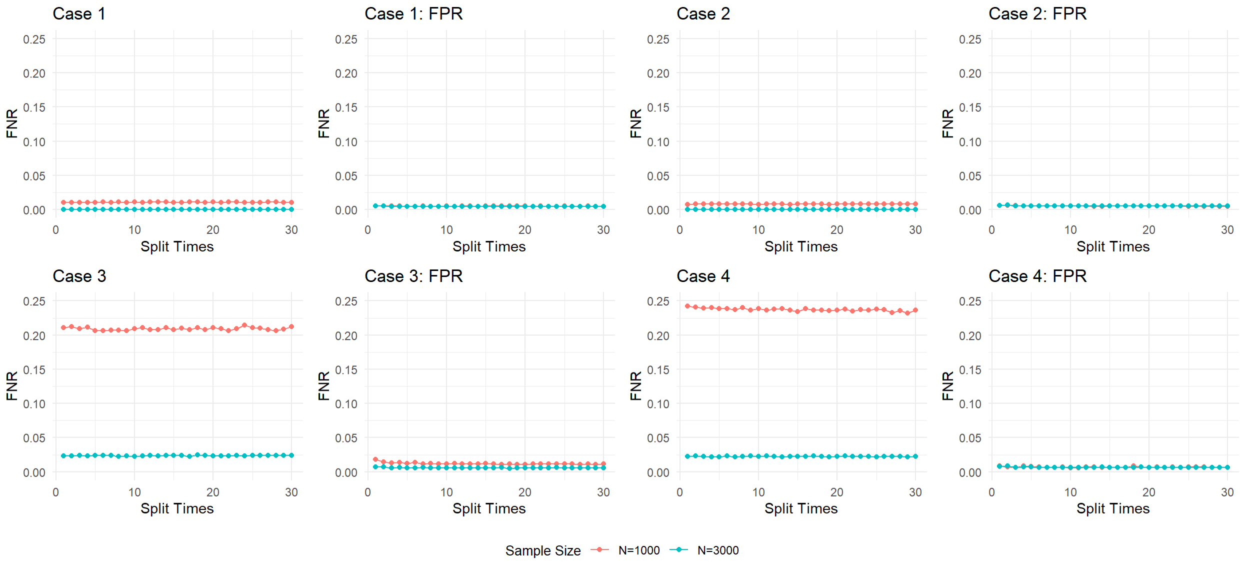


Figure 3.4 False positive rate and false negative rate as the number of sample splits changes from 0 to 30.

The error terms ε_x, ε_y were generated from $(\varepsilon_{xi}, \varepsilon_{yi}) \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{pmatrix}$, $i = 1, ..., N$. For the effect size of α and γ, we considered the following four cases:

- Case 1: $\alpha = (\underbrace{0.4, \cdots, 0.4}_{21})$, $\gamma = (\underbrace{0, \cdots, 0}_{9}, \underbrace{0.4, \cdots, 0.4}_{6}, \underbrace{0.2, \cdots, 0.2}_{6})$
- Case 2: $\alpha = (\underbrace{0.15, \cdots, 0.15}_{21})$, $\gamma = (\underbrace{0, \cdots, 0}_{9}, \underbrace{0.4, \cdots, 0.4}_{6}, \underbrace{0.2, \cdots, 0.2}_{6})$
- Case 3: $\alpha = (\underbrace{0.15, \cdots, 0.15}_{5}, \underbrace{0.07, \cdots, 0.07}_{16})$,

  $$\gamma = (0, 0, 0, 0.2, 0.1, \underbrace{0, \cdots, 0}_{6}, \underbrace{0.2, \cdots, 0.2}_{5}, \underbrace{0.1, \cdots, 0.1}_{5})$$
- Case 4: $\alpha = (\underbrace{0.07, \cdots, 0.07}_{21})$, $\gamma = (\underbrace{0, \cdots, 0}_{9}, \underbrace{0.2, \cdots, 0.2}_{6}, \underbrace{0.1, \cdots, 0.1}_{6})$

Case 1 and Case 2 came from the simulation settings of WIT method (Lin et al., 2024). Case 3 and Case 4 examined scenarios where the effects of IVs on X and Y were weaker compared to Case 1 and Case 2. In Case 3, the effects of IVs on X were heterogeneous, with a small subset of IVs exhibiting stronger associations. Invalid IVs were evenly distributed across these two groups of

IVs. Case 4, on the other hand, is considered an even weaker IV scenario. This adjustment stems from the fact that the situations originally considered in WIT, when measured by the first-stage F-statistics, are still far from the commonly considered weak IV threshold (i.e., $F < 10$) (Staiger and Stock, 1997). Specifically, when $N = 1000$, the F-statistic is 513 in Case 1 and 73 in Case 2 (see Table 3.1). Therefore, there remains room to further lower the F-statistics. To this end, we considered more extreme scenarios in Case 3 and Case 4, which present great challenges. In addition, since this simulation setting assumes no noise, there is no selection bias to mitigate. Therefore, we performed only 10 sample splits, which is sufficient in this case.

Table 3.1 Average F-statistics for the first stage of 2SLS in simulations without noise.

	N=1000	N=3000
Case 1	513.65	1536.61
Case 2	73.06	216.85
Case 3	33.02	96.90
Case 4	16.68	47.97

Figure 3.5 shows the violin plots for the estimates obtained from each method under different cases and sample sizes. Firstly, it is evident that the sisVIVE method exhibits a significantly larger estimation bias compared to other methods across all cases and sample sizes. This is because sisVIVE relies on a simple LASSO regression to estimate α and assumes that more than half of the IVs are valid. However, the scenarios we considered violate this assumption. In Case 1 and Case 2, although CIIV and WIT appear to produce estimates clustered around the true value, they also generate some extreme outliers that deviate significantly from the true value. Comparatively, CIIV performs slightly better than WIT under strong IV and large sample conditions. This is because CIIV classifies IVs by constructing confidence intervals for the causal effect using each SNP individually as an IV. With strong IVs and large samples, the constructed confidence intervals are more reliable. However, this approach becomes less effective in scenarios with weak IVs. As a result, in Case 3 and Case 4, the bias of CIIV's estimates progressively increases. While WIT performs slightly better than CIIV when IVs are weak or sample sizes are small, it still produces many estimates that deviate substantially from the true value. Furthermore, as the sample size increases, its performance

is even worse than that of CIIV. On the other hand, MR-SPLIT+ consistently produces results that closely match those of the oracle TSLS estimates across all conditions, regardless of whether the IVs are strong or weak, under different sample sizes.

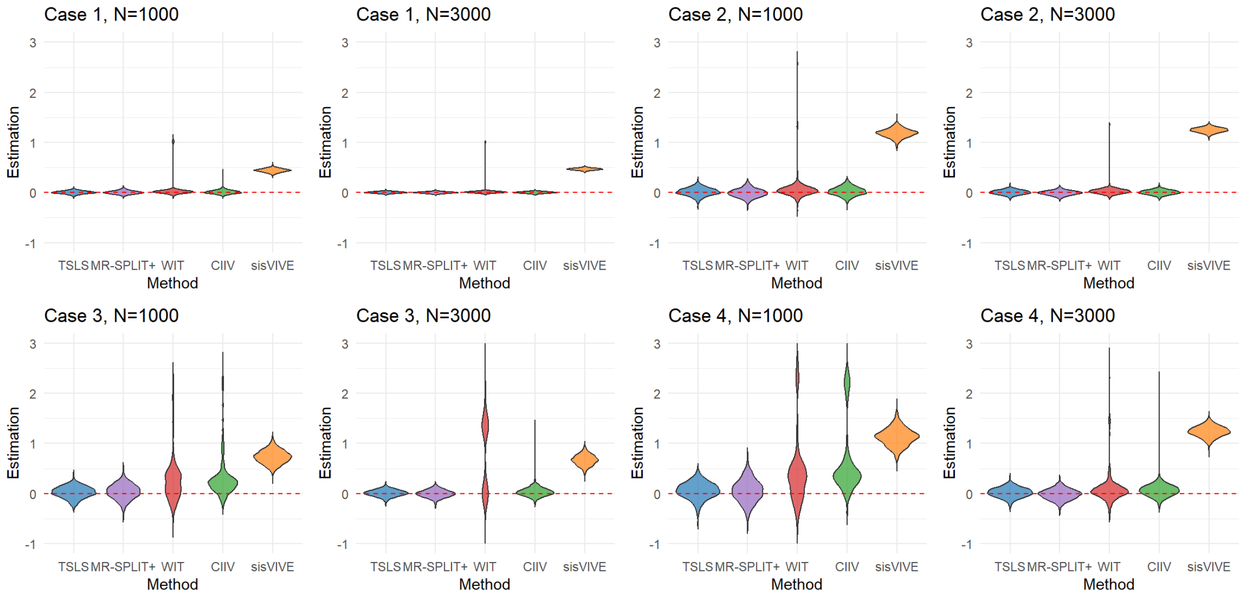
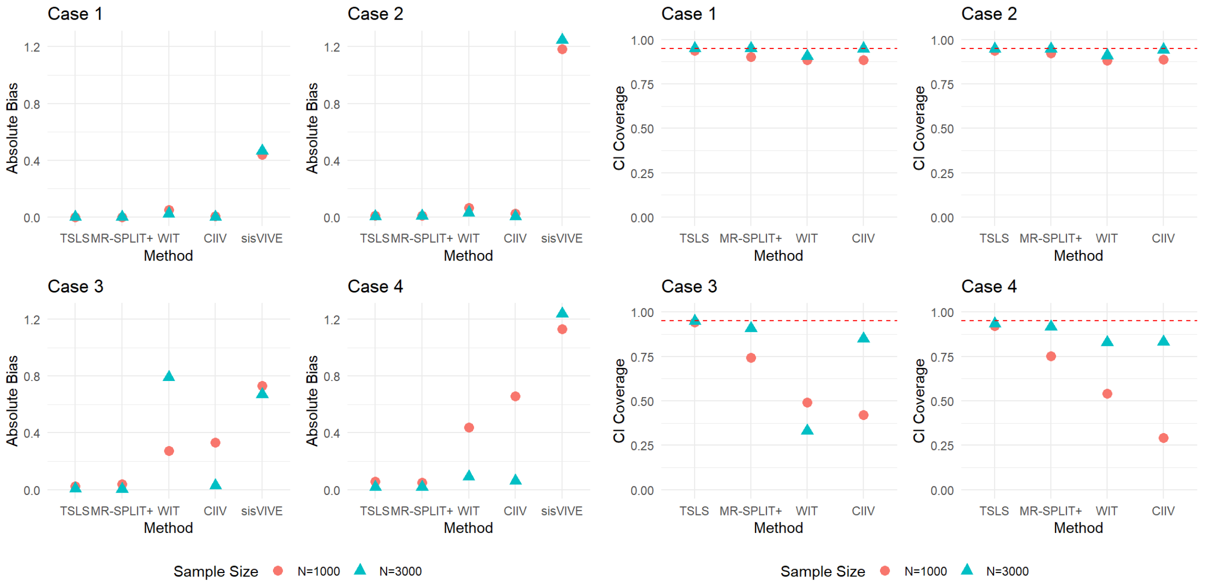


Figure 3.5 Violin plots of the causal estimates in simulations without noise.

Figure 3.6a shows the absolute bias of the estimates obtained by these methods across various cases. Notably, the bias produced by MR-SPLIT+ is almost as small as that of the oracle TSLS estimates, while all other methods exhibit significantly larger biases. CIIV, however, only achieves comparable results when the sample size is sufficiently large. Figure 3.6b presents the coverage probabilities obtained by each method. sisVIVE is excluded from this analysis as it does not provide a way to construct the confidence interval. When the sample size is large ($N = 3000$), MR-SPLIT+ achieves nearly 95% coverage probability, even under scenarios with weak IVs (Case 3 and Case 4). For a smaller sample (e.g., $N = 1000$), although MR-SPLIT+ does not reach the 95% coverage probability, it consistently demonstrates superior performance compared to WIT and CIIV across all cases.

Figure 3.7 illustrates the False Negative Rate (FNR) and False Positive Rate (FPR) for identifying invalid IVs. The FNR represents the proportion of invalid IVs incorrectly identified as valid ones, while the FPR is the proportion of valid IVs incorrectly identified as invalid ones. Controlling the



(a) Absolute bias of estimators.

(b) Coverage Probability.

Figure 3.6 Comparison of absolute bias (a) and coverage probability (b) for different methods in simulations without noise.

FNR is crucial because misclassifying invalid IVs as valid ones directly introduces systematic bias into the causal effect estimates, potentially leading to severely biased conclusions. In contrast, a higher FPR, which results in the unnecessary exclusion of valid IVs, primarily affects the efficiency of the estimation rather than introducing substantial bias. The results demonstrate that MR-SPLIT+ consistently achieves accurate identification of invalid IVs, maintaining a very low FPR rate. In contrast, sisVIVE produces significantly higher FPR values followed by WIT. Regarding the FNR, MR-SPLIT+ achieves values comparable to those of WIT, CIIV, and sisVIVE under small sample sizes ($N = 1000$), while demonstrating substantially lower FNR under larger sample sizes ($N = 3000$).

We also presented the actual breakdown of the selected valid and invalid IVs in Tables A.2 and Table A.3, respectively. These tables were used to compute the FPR and FNR. When the IVs are strong, as in Case 1 and Case 2, MR-SPLIT+ demonstrated exceptionally high accuracy in distinguishing between valid and invalid IVs. For instance, in Case 1 with a sample size of $N = 1000$, MR-SPLIT+ identified an average of 9 valid IVs and 0.1 invalid ones as valid IVs while the true number of valid IVs was 9 in each simulation run, which means it correctly selected all

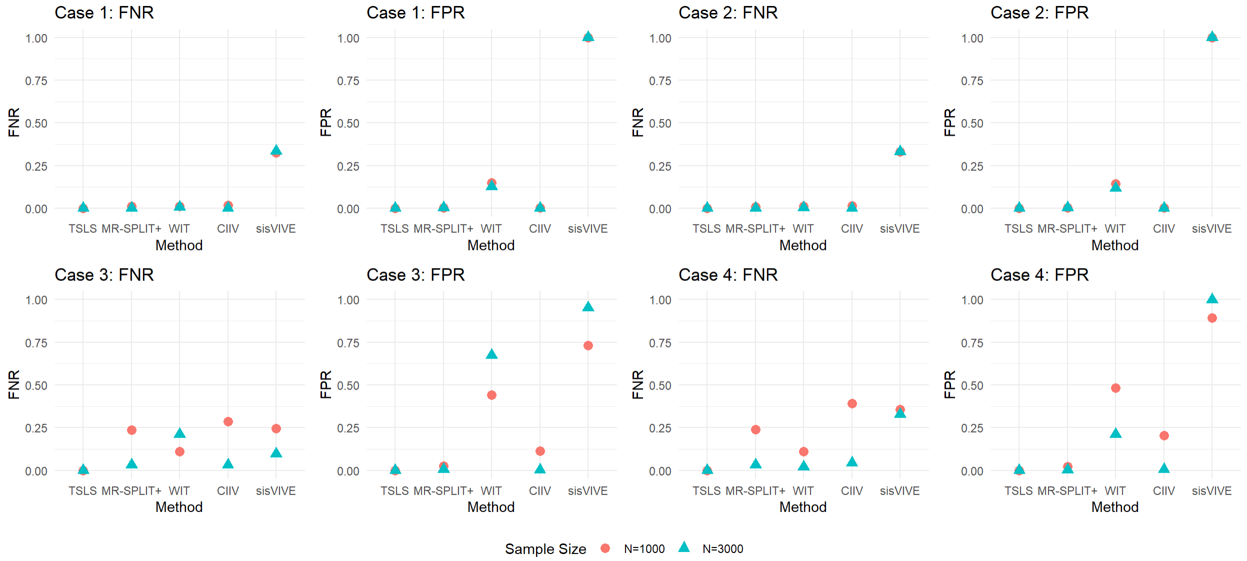


Figure 3.7 Plots of FNR and FPR for IV selection in simulations without noise.

the valid ones and only 0.1 out of 9.1 IVs were misidentified (see Table A.2). In comparison, WIT identified 7.7 valid and 0.2 invalid IVs as valid IVs. CIIV performed similarly to MR-SPLIT+, selecting 9 valid and 0.2 invalid IVs as valid IVs. However, in Case 3 and Case 4, the performance of CIIV deteriorated considerably, particularly when $N = 1000$. Regarding the classification of invalid IVs (see Table A.3), MR-SPLIT+ showed even stronger performance, with most cases involving no misclassification of valid IVs as invalid. In contrast, WIT consistently misclassified a substantial proportion of valid IVs as invalid, indicating a tendency toward over-selection of invalid instruments.

3.4.3 Simulation with noise IVs

To better mimic real world conditions, we also explored a more realistic scenario by selecting IVs from a pool of candidates that included both relevant instruments and noise variables. Specifically, we generated a total of 300 variables and randomly picked 21 to have effects on the exposure X. The remaining 279 variables are noise ones. The other settings remained consistent with those described in Section 4 in the manuscript. To ensure a fair comparison, all methods (except the oracle TSLS method) used the first stage of CIIV to select relevant IVs for the exposure. Specifically, MR-SPLIT+ performed selection within each subset, while other methods conducted selection on the whole sample. To ensure the robustness of MR-SPLIT+ estimations, the sample was split

30 times in each scenario. Additionally, all methods were evaluated through 1000 replications to comprehensively assess their performance.

Figure A.25 in the Appendix presents violin plots of the estimates obtained by each method under various cases and sample sizes in the presence of noise IVs. The results are generally consistent with those observed in the absence of noise IVs. For the WIT, CIIV, and sisVIVE methods, their estimates exhibit larger bias compared to scenarios without noise IVs. In contrast, MR-SPLIT+ maintains nearly the same level of superior performance, demonstrating robustness to the presence of noise IVs. This observation is further supported by the results shown in Figures A.26 and A.27.

Figure A.27 illustrates the FNR and FPR for identifying invalid IVs. The results demonstrate that in the presence of noise IVs, the performance of MR-SPLIT+ remains comparable to its performance in the absence of noise, highlighting its superior ability to mitigate the selection bias issue. In contrast, other methods exhibit a decline in performance compared to noise-free scenarios, with the deterioration being particularly pronounced in small sample sizes ($N = 1000$).

The actual breakdown used to compute FNR and FPR shown in Figure A.27 is presented in Tables A.4 and A.5 in the Appendix. Overall, MR-SPLIT+ maintains the highest accuracy of correctly identifying those true valid IVs and of avoiding the incorrect classification of invalid or noise IVs as valid. For example, in Table A.4, in Case 3 with a sample size of $N = 3000$, MR-SPLIT+ identified an average of 8.9 valid IVs out of 9 true IVs, while 0.3 invalid IVs and 1.4 noise were misidentified as valid IVs. In comparison, WIT identified only 4.4 true valid IVs out of 9 with a higher misclassification rate for invalid (1.7) and noise IVs (0.2). In all the cases, sisVIVE has the worst performance.

3.5 Real Data Application

We evaluated two datasets described in Section 3.2. MR-SPLIT+ was evaluated alongside TSLS, sisVIVE, CIIV, and WIT. The results obtained from MR-SPLIT+ are presented together with those from the other methods to facilitate a comprehensive performance comparison.

3.5.1 Positive control: casual effect of BMI on DBP

The dataset initially comprised 502,505 individuals and included a total of four instances. Considering both the accuracy of the dataset and the size of the sample, we selected version 1.0 for our analysis, which was conducted between 2012-2013. We filtered the data to include only samples with complete exposure and outcome information, reducing the sample size to 50,497. After merging this dataset with genetic data, we obtained a final sample of 39,889 individuals. Additionally, we calculated the age of participants at the time of measurement based on their 'birth year' and their 'date attending assessment center'. Among them, 12,022 individuals were aged between 44 and 59 years. The entire population ranged from 44 to 83 years. Considering the impact of age on blood pressure, we performed MR analyses separately for the 44–59 age group ('young group') and the overall population ('all group').

For the genetic data, we excluded variants with a missing rate above 10% and a MAF below 0.05. Additionally, LD pruning was applied to remove highly correlated variants, a common practice in MR analysis. This preprocessing resulted in a final set of 279k SNPs. We initially referred to the GWAS study on BMI conducted by Locke et al. (Locke et al., 2015), which identified 97 variants with p-values $< 5 \times 10^{-8}$. After matching these variants with our existing genetic data, we identified 21 candidate IVs. However, their associations with BMI in our sample were weak, with some even exhibiting p-values greater than 0.05. Thus, we applied the SIS method (Fan and Lv, 2008) to select the top 80 SNPs mostly associated with BMI as candidate IVs separately for the 'young group' and the 'all group'. After combining the candidate IVs identified earlier from the GWAS study, we obtained 101 candidate IVs for the 'young group' and 99 candidate IVs for the 'all group', respectively.

Table 3.2 shows the results of the 'young group'. In the simplest case of OLS regression, the two variables demonstrated a very strong correlation (p-value $< 2 \times 10^{-16}$). When estimating the causal relationship using various MR methods, all approaches yielded p-values below 0.05. Notably, CIIV and MR-SPLIT+ performed additional filtering of the candidate IVs, removing those that were potentially too weak or noisy. MR-SPLIT+ retained 46 relevant IVs, while CIIV retained

69. For reference, the F-statistic in the first stage of TSLS was 12.37, indicating that this represents a relatively weak set of IVs, which could lead to potential weak instrument bias.

Table 3.2 Comparison of results in the 'young group' when assessing the causal effect of log(BMI) on log(DBP).

Method	$\hat{\beta}$	**s.d.**	**pvalue**	**lower CI**	**upper CI**	**Relevant IVs**	**Valid IVs**	**Invalid IVs**
OLS	0.2814	0.0067	<1e-15	0.2683	0.2945	NA	NA	NA
TSLS	0.2690	0.0231	<1e-15	0.2238	0.3142	101	101	0
sisVIVE	0.2690	NA	NA	NA	NA	101	101	0
CIIV	0.2792	0.0236	<1e-15	0.2329	0.3255	69	69	0
WIT	0.2675	0.0238	<1e-15	0.2208	0.3142	101	101	0
MR-SPLIT+	0.2589	0.0551	2.7e-06	0.1509	0.3669	46	46	0

Note: Sample size $N = 12,022$, Age: $44 \sim 59$, F-statistic = 12.37 in the first stage of TSLS..

Table 3.3 shows the results of the 'all group'. Consistent with previous findings, significant results were obtained across all MR methods when considering the entire population. Among these, MR-SPLIT+ identified 56 relevant IVs from the 99 candidate ones and filtered 1 invalid IV. The estimated causal effect was $\hat{\beta} = 0.2133$, which is slightly lower than the causal effect estimated in the 'young group'.

Table 3.3 Comparison of results in the 'all group' when assessing the causal effect of log(BMI) on log(DBP).

Method	$\hat{\beta}$	**s.d.**	**pvalue**	**lower CI**	**upper CI**	**Relevant IVs**	**Valid IVs**	**Invalid IVs**
OLS	0.2249	0.0040	<2e-16	0.2170	0.2328	NA	NA	NA
TSLS	0.1972	0.0221	<2e-16	0.1540	0.2404	99	99	0
sisVIVE	0.1972	NA	NA	NA	NA	99	99	0
CIIV	0.2186	0.0244	<2e-16	0.1708	0.2664	65	64	1
WIT	0.0794	0.0331	1.7e-02	0.0145	0.1444	99	48	51
MR-SPLIT+	0.2133	0.0400	9.6e-08	0.1349	0.2916	56	55	1

Note: Sample size $N = 39,889$, Age: $44 \sim 83$, F-statistic = 15.03 in the first stage of TSLS.

Following the significant causal effect of BMI on DBP, we conducted downstream mediation analyses to explore potential biological pathways that may underlie this relationship. Candidate mediators were selected based on prior biological knowledge linking adiposity to blood pressure regulation. Specifically, we focused on biomarkers representing metabolic, inflammatory, renal, hematologic, and lipid-related processes. To avoid redundancy and ensure interpretability, we excluded variables that are strongly collinear with BMI (e.g., waist circumference, hip circumference,

Table 3.4 Results of mediation analysis between log(BMI) and log(DBP).

Mediator	$\hat{\alpha}$	$\hat{\beta}$	Prop	adj. p-value
RBC	0.5070	0.0620	0.1392	< 0.0001
log(CRP)	2.5321	0.0033	0.0367	0.0165
log(Creatinine)	0.2205	0.0208	0.0201	0.0003
log(Glucose)	0.1556	0.0190	0.0131	0.0956
HDL	-0.8902	0.0058	0.0228	0.5142

Note: $\hat{\alpha}$ denotes the estimated effect from log(BMI) to the mediator, and $\hat{\beta}$ denotes the estimated effect from the mediator to log(DBP), controlling for log(BMI). Prop represents the proportion of the total effect that is mediated through the mediator.

and regional fat mass) or lacked clear biological plausibility as mediators. The final set of mediators included red blood cell count (RBC), C-reactive protein (CRP), serum creatinine, fasting glucose, and high-density lipoprotein (HDL). These variables were retained due to their well-established physiological relevance in the context of obesity and cardiovascular regulation. Specifically, RBC reflects hematologic changes that may influence blood viscosity and vascular resistance; CRP is a widely used marker of systemic inflammation; creatinine serves as an indicator of renal function, which is closely linked to blood pressure control; glucose levels capture metabolic disturbances associated with insulin resistance and sympathetic activation; and HDL represents lipid metabolism, which has been implicated in the development of hypertension. These mediators collectively reflect diverse biological domains through which BMI may exert its influence on diastolic blood pressure. The analysis was conducted on all available samples at version 1.0. Given the wide age range, age was included as a covariate in the model. The significance of each candidate mediator was assessed using the traditional Sobel test (Sobel, 1982), with Bonferroni correction applied for multiple testing.

Table 3.4 presents the results of the mediation analysis assessing the biological pathways through which BMI influences DBP. We also constructed a Directed Acyclic Graph (DAG), see Fig 3.8, to illustrate the variable relationships, including only the statistically significant mediators. Among the selected mediators, red blood cell count (RBC) mediated the largest proportion of the effect (13.9%), consistent with the role of obesity in stimulating erythropoiesis, increasing blood viscosity, and contributing to vascular resistance (He et al., 2023). C-reactive protein (CRP) and serum creatinine were also significant mediators, contributing 3.7% and 2.0% of the total effect, respectively. These findings support the hypothesis that systemic inflammation and early renal function impairment are important biological mechanisms underlying obesity-associated increases in diastolic blood pressure (Coresh et al., 2001). Although glucose and HDL were included based on their established metabolic and lipid-related roles, their mediation proportions were relatively modest (1.3% and 2.2%, respectively), and the associations did not reach statistical significance after multiple testing correction.

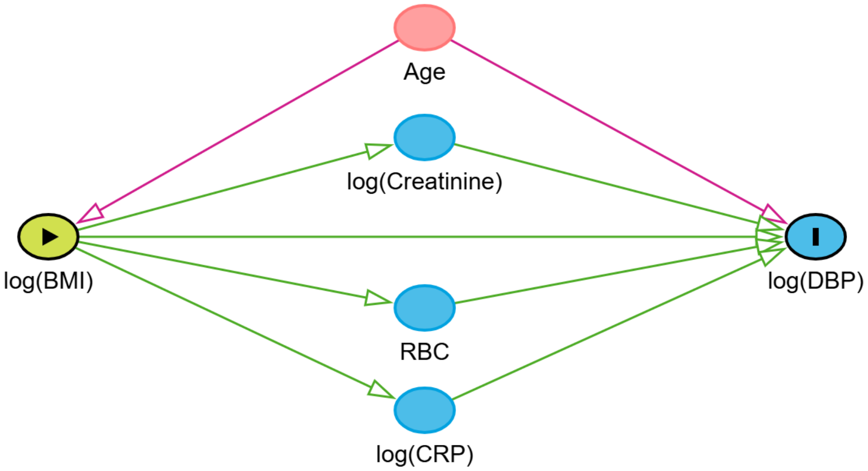


Figure 3.8 DAG representing the significant mediation pathways between BMI and DBP.

Overall, these results underscore the multifactorial nature of the BMI–DBP relationship, implicating hematologic, inflammatory, and renal pathways as key contributors.

3.5.2 Negative control: no causal effect of BMI on BW

As in the previous analysis, the initial sample size was 502,505, and we selected version 0.0 for analysis based on the dataset size, which covers data collected between 2006 and 2010. To more effectively evaluate the robustness of our method, we focused on the age group with the strongest association, identified as individuals aged 60 to 71 years. After merging this subset with the genetic data, the sample size was reduced to 85,248.

OLS regression revealed a significant correlation between log(BMI) and BW, with a p-value less than 2×10^{-16}. For candidate IV selection, similar to the previous approach, we first identified 21 SNPs associated with BMI from the GWAS study (Locke et al., 2015) within our genetic dataset. Subsequently, we selected an additional 80 top SNPs strongly associated with BMI from the dataset to serve as candidate IVs. Together, we had 101 SNP IVs included in the analysis. Table 3.5 presents the causal effect estimates obtained using different MR methods. P-values highlighted in bold font indicate statistically significant results, which, in this context, indicate a false positive. Specifically, the TSLS method reported an estimate of $\hat{\beta} = 0.2279$ with a p-value of 0.0413, suggesting that treating all selected candidate IVs as valid IVs can lead to biased estimates. Similarly, the WIT method yields an even more biased estimate of $\hat{\beta} = -0.49$ with a p-value of 0.044. This bias may stem from the method itself identifying too many invalid IVs, which are likely misclassified, especially when compared to CIIV and MR-SPLIT+. In contrast, both CIIV and MR-SPLIT+ produce non-significant results and estimates closer to zero, indicating greater reliability and alignment with the expected null causal effect compared to the other methods.

Table 3.5 Comparison of results when assessing the causal effect of log(BMI) on birth weight.

Method	$\hat{\beta}$	**s.d.**	**pvalue**	**lower CI**	**upper CI**	**Relevant IVs**	**Valid IVs**	**Invalid IVs**
OLS	0.2193	0.0150	<2e-16	0.1900	0.2486	NA	NA	NA
TSLS	0.2279	0.1117	0.0413	0.0089	0.4468	100	100	0
sisVIVE	0.2279	NA	NA	NA	NA	100	100	0
CIIV	0.0382	0.1542	0.8044	-0.2640	0.3403	35	35	0
WIT	-0.4900	0.2433	0.0440	-0.9668	-0.0132	100	23	77
MR-SPLIT+	0.1279	0.1798	0.4767	-0.2244	0.4803	35	35	0

Note: Sample size $N = 85,248$, Age: $60 \sim 71$, F-statistic = 15.59 in the first stage of TSLS.

3.6 Discussion

In this chapter, we proposed MR-SPLIT+, an innovative extension of the MR-SPLIT method in one-sample MR analysis, under the relaxed plurality rule. Building upon the ability to address selection bias and weak IV issues, MR-SPLIT+ further allows for the handling of invalid IVs, one of the great challenges in MR analysis. The incorporation of the best subset selection method and multiple splitting techniques enhances the robustness of the approach, significantly improving the accuracy of invalid IV identification. This feature makes MR-SPLIT+ a powerful and reliable tool for one-sample MR analysis. The result of the selection consistency provides a theoretical guarantee for the method.

In fact, the assumptions required by MR-SPLIT+ can be further relaxed, and that our currently proposed version of the relaxed plurality rule may still be overly restrictive. Due to limitations in current theoretical development, we are unable to formally establish results for the case where the specified number of invalid IVs k is smaller than the true number of invalid IVs $|A_0|$. However, in practice, our method often tends to prioritize the selection of invalid IVs under such misspecification. In other words, valid IVs are more likely to be shrunk toward zero, and thus excluded from the model. Moreover, since the selected set of valid IVs must pass an overidentification test to confirm that it contains only one group of instruments, choosing both valid IVs and invalid IVs (i.e., setting $p - k > p - |A_0|$) often leads to failure in this test. Consequently, the algorithm tends to increment k step by step until the true value is reached. This iterative process implies that, under certain conditions, our method may still work even when the number of valid IVs is smaller than that of the invalid ones. Although we are currently unable to provide formal theoretical guarantees for this behavior, empirical evidence supports our belief that the proposed method has strong robustness and broad applicability in practical settings.

We conducted simulation studies to compare MR-SPLIT+ with state-of-the-art methods proposed in recent years. The results demonstrate that, regardless of whether IVs are strong or weak, our method consistently outperforms others, often achieving results comparable to the oracle TSLS method (assuming the true set of valid IVs is known in advance). The findings from the multiple splitting procedure further underscore its necessity, as the estimators obtained through repeated splits yield more robust estimates and coverage probabilities closer to the nominal 95% level.

Though MR-SPLIT+ involves multiple sample splits, it does not necessarily introduce a high computational cost compared to other approaches. For example, in the analysis of the BMI and DBP dataset, the ‘young group’ consists of 12,022 samples. MR-SPLIT+ required only 1.39 seconds per split, and 30 splits took only 41.7 seconds. In contrast, WIT took 177.64 seconds, and sisVIVE required 102.00 seconds. When the sample size increased to 39,889 in the ‘all group’, MR-SPLIT+ took approximately 9.94 seconds per split, and 30 splits required only 298.3 seconds. In comparison, WIT required 661.91 seconds. Furthermore, as the sample size increases, researchers can opt to reduce the number of splits in MR-SPLIT+ accordingly, as shown in the simulation study, further saving the computational cost.

While the current method is already well-developed and robust, there remains room for further refinement to broaden its applicability to more diverse scenarios, which will be investigated in our future work. For instance, it could be adapted to accommodate binary outcomes and binary exposures, a task that should be relatively straightforward. Additionally, the method could be expanded to address bidirectional causal inference, enabling researchers to study reciprocal causal relationships more effectively. Furthermore, an exciting direction for future research lies in extending MR-SPLIT+ to construct causal networks involving multiple exposures, facilitating a more comprehensive understanding of complex causal structures in human diseases.

In summary, MR-SPLIT+ holds immense potential for further development and applications, making it a versatile and powerful tool for advancing causal inference research.

CHAPTER 4

BIMR-SPLIT+ — BIDIRECTIONAL MR AND CAUSAL MECHANISM

4.1 Introduction

Understanding bidirectional or ambiguous causal relationships is essential in many scientific domains, such as epidemiology, economics, and social sciences. In practice, it is common to encounter situations where either two traits may influence each other, or the direction of causality is unknown. For example, the relationship between physical activity and mental health (Schuch et al., 2018; Mammen and Faulkner, 2013), or between inflammation and depression (Khandaker et al., 2014), may involve feedback loops or unclear temporal precedence. A particularly important application arises in the construction of gene regulatory networks (Albert and Kruglyak, 2015), where distinguishing between causal gene expression (which influences disease risk) and response gene expression (which is influenced by the disease) is critical. Accurately identifying causal genes enables researchers to prioritize therapeutic targets and avoid misdirected interventions that focus on downstream biomarkers rather than the true drivers of disease. In such cases, robust statistical methods that can infer or test for potential bidirectional causality are crucial for valid scientific conclusions and effective policy or intervention design.

A typical approach in MR studies addressing potential bidirectional causality is to simply apply univariable MR analyses in both directions—treating one trait as the exposure and the other as the outcome, and then reversing the roles (Davey Smith and Hemani, 2014; Zhao et al., 2023; Maina et al., 2023). However, this naive strategy often overlooks critical assumptions of MR, especially the validity of the IVs in both directions. When the same set of genetic variants influences both traits or when pleiotropy is present, the core IV assumptions may be violated, leading to biased and misleading causal estimates.

In unidirectional Mendelian Randomization, numerous methods have been developed in recent years to address the issue of invalid IVs. Notable examples include MR-Egger (Bowden et al., 2015), sisVIVE (Kang et al., 2016), CIIV (Windmeijer et al., 2021), and WIT (Windmeijer et al., 2021), each of which relies on specific identifying assumptions. Among them, the plurality rule assumed by CIIV is relatively mild and has been considered advantageous in practice. However, this assumption is inherently violated in the presence of bidirectional causality. Consider a setting where the exposure and outcome exert causal effects on each other. In such bidirectional scenarios, the validity of CIIV's plurality rule is compromised. Specifically, when the number of SNPs affecting the outcome exceeds the number of SNPs affecting the exposure, the instruments that primarily influence the outcome may be mistakenly selected as valid IVs for estimating the causal effect from exposure to outcome. This misclassification arises because the outcome, in turn, influences the exposure, thereby inducing reverse associations that distort the instrument strength ranking required by CIIV. As a result, the plurality rule, which assumes that the largest group of instruments reflects the true causal direction, no longer holds. We will provide a detailed discussion of this issue in a later section.

MR-SPLIT+ (Shi et al., 2025) is a recently proposed unified framework designed to address several key challenges in one sample MR, including selection bias, weak instruments, and the presence of invalid IVs. Notably, the assumptions underlying MR-SPLIT+ are even more relaxed than the commonly adopted plurality rule, thereby offering a promising solution in settings with bidirectional causality, where traditional assumptions often fail. Furthermore, the methodological structure of MR-SPLIT+ is closely aligned with that of TSLS (Angrist et al., 1996), allowing for considerable flexibility and adaptability in implementation. This combination of robustness and flexibility makes MR-SPLIT+ a valuable tool for causal inference in complex MR scenarios.

In this study, inspired by the work Chen (2025), we extend the MR-SPLIT+ framework to the context of bidirectional MR and named it as BiMR-SPLIT+. Leveraging the inherent flexibility of the original model, we introduced several methodological modifications that substantially improved its computational efficiency. To rigorously assess the performance of the proposed approach, we conducted extensive simulation studies under a wide range of realistic scenarios. Furthermore, we generalized the method to the construction of causal networks, aiming to capture the mutual influences between gene expression and complex traits. Due to the challenges associated with obtaining large-scale human datasets, we applied our approach to a dataset of approximately 180

Drosophila melanogaster individuals, focusing on uncovering bidirectional causal relationships between gene expression levels and phototactic behavior.

4.2 Model and Methodology

Suppose we are interested in the causal effects between $X \in \mathbb{R}^{N\times 1}$ and $Y \in \mathbb{R}^{N\times 1}$, consider the following models:

$$\begin{aligned} U &= G\alpha_u + \varepsilon_u \\ X &= G\alpha_x + Y\beta_{YX} + U\eta_x + \varepsilon_x, \\ Y &= G\alpha_y + X\beta_{XY} + U\eta_y + \varepsilon_y \end{aligned} \tag{4.1}$$

where $U \in \mathbb{R}^{N\times p_u}$ represents the unobserved confounders that may affect both X and Y. The parameters β_{XY} and β_{YX} denote the causal effects of interest from X to Y and from Y to X, respectively. $G \in \mathbb{R}^{N\times p}$ denotes the matrix of SNPs that may influence X, Y, or both. The error terms ε_u, ε_x, and ε_y are assumed to follow independent normal distributions. For the j-th SNP, G_j is considered invalid when estimating β_{XY} if $\alpha_y + \alpha_u\eta_y \neq 0$, and invalid when estimating β_{YX} if $\alpha_x + \alpha_u\eta_x \neq 0$. See Figure 4.1 for an illustration.

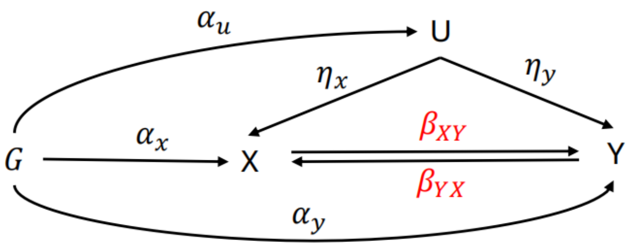


Figure 4.1 Bidirectional MR.

To simplify, we rewrite Equations 4.1 as follows:

$$\begin{aligned} X &= G(\alpha_x + \alpha_u\eta_x) + Y\beta_{YX} + (\varepsilon_x + \varepsilon_u\eta_x) \\ &= G\alpha_1 + Y\beta_{YX} + \varepsilon_1, \\ Y &= G(\alpha_y + \alpha_u\eta_y) + X\beta_{XY} + (\varepsilon_y + \varepsilon_u\eta_y) \\ &= G\alpha_2 + X\beta_{XY} + \varepsilon_2. \end{aligned} \tag{4.2}$$

where $\alpha_1 = \alpha_x + \alpha_u\eta_x$ and $\alpha_2 = \alpha_y + \alpha_u\eta_y$. In this setting, G_j is considered invalid when estimating β_{XY} if $\alpha_2 \neq 0$, and invalid when estimating β_{YX} if $\alpha_1 \neq 0$. The error terms ε_1 and

ε_2 are assumed to follow a bivariate normal distribution with nonzero covariance, induced by the presence of unobserved confounders affecting both X and Y.

4.2.1 Stage one

Based on the values of α_{1j} and α_{2j}, we classify SNP j into one of the following three categories:

- $S_X = \{j : \alpha_{1j} \neq 0, \alpha_{2j} = 0\}$: valid for X;
- $S_Y = \{j : \alpha_{1j} = 0, \alpha_{2j} \neq 0$: valid for Y;
- $S_I = \{j : \alpha_{1j} \neq 0, \alpha_{2j} \neq 0$: invalid for both X and Y.

In the following, we use G_A to denote the submatrix of G consisting of SNPs indexed by the set A, i.e., $G_A = \{G_j : j \in A\}$. Note that SNPs from all three groups may be selected as relevant instruments for either X or Y, as the bidirectional causal relationship between X and Y can induce correlations between G_j and both traits, regardless of the true direction of validity.

Nevertheless, while SNPs in $G_{S_Y \cup S_I}$ are invalid instruments for estimating the causal effect from X to Y, those that are also relevant for X can still be included as covariates to reduce variance and enhance estimation efficiency. Therefore, in stage one of the MR-SPLIT+ procedure, we recommend using all selected relevant IVs without excluding potentially invalid ones.

For each direction, we first split the sample evenly into two equally sized subsets, then perform IV selection separately in both subsets and take the union of the selected IV sets. Let $\hat{S}_{X1}$ denote the union of selected IVs for X, and let $\hat{X}$ be the corresponding fitted value. Similarly, let $\hat{S}_{Y1}$ denote the union of selected IVs for Y, and let $\hat{Y}$ be the corresponding fitted value.

4.2.2 Stage two

Following the work of MR-SPLIT+, we identify invalid IVs for X by solving:

$$\hat{\alpha}_{2,\hat{S}_{X1}} = \arg\min_{\alpha_{2,\hat{S}_{X1}}} \|\alpha_{2,\hat{S}_{X1}}\|_0 \quad \text{s.t.} \quad \|Y - \widetilde{G}_{\hat{S}_{X1}}\alpha_{2,\hat{S}_{X1}}\|_2^2 < \delta, \tag{4.3}$$

where $\widetilde{G}_{\hat{S}_{X1}} = M_{\hat{X}} G_{\hat{S}_{X1}} = (I - \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}')G_{\hat{S}_{X1}}$.

Similarly, for the direction from Y to X, we identify invalid IVs for Y by solving:

$$\hat{\alpha}_{1,\hat{S}_{Y1}} = \arg\min_{\alpha_{1,\hat{S}_{Y1}}} \|\alpha_{1,\hat{S}_{Y1}}\|_0 \quad \text{s.t.} \quad \|X - \widetilde{G}_{\hat{S}_{Y1}}\alpha_{1,\hat{S}_{Y1}}\|_2^2 < \delta, \tag{4.4}$$

where $\widetilde{G}_{\hat{S}_{Y1}} = M_{\hat{Y}} G_{\hat{S}_{Y1}} = (I - \hat{Y}(\hat{Y}'\hat{Y})^{-1}\hat{Y}')G_{\hat{S}_{Y1}}$.

Before directly applying the stage two procedure of MR-SPLIT+ to identify invalid IVs, we first exploit an observation that allows for the rapid pre-screening of a subset of clearly invalid instruments.

From Equation 4.2, we can derive the following inequalities, as shown in Chen (2025):

$$\begin{aligned} \mathrm{Var}(X) &> \beta_{YX}^2 \,\mathrm{Var}(Y), \\ \mathrm{Var}(Y) &> \beta_{XY}^2 \,\mathrm{Var}(X). \end{aligned} \tag{4.5}$$

By substituting the first inequality into the second, we obtain:

$$\mathrm{Var}(Y) > \beta_{XY}^2 \,\mathrm{Var}(X) > \beta_{XY}^2 \beta_{YX}^2 \,\mathrm{Var}(Y),$$

which implies that $\beta_{XY}\beta_{YX} < 1$.

This result holds under the assumption that the causal effects β_{XY} and β_{YX} are well-defined and finite. The condition $\beta_{XY}\beta_{YX} < 1$ further suggests that a bidirectional feedback system between X and Y cannot exhibit unbounded amplification.

We could also rewrite Equation 4.1 as followings:

$$\begin{aligned} X &= \frac{1}{1 - \beta_{YX}\beta_{XY}} [G(\alpha_2 \beta_{YX} + \alpha_1) + (\varepsilon_2 \beta_{YX} + \varepsilon_1)], \\ Y &= \frac{1}{1 - \beta_{YX}\beta_{XY}} [G(\alpha_1 \beta_{XY} + \alpha_2) + (\varepsilon_1 \beta_{XY} + \varepsilon_2)] \end{aligned} \tag{4.6}$$

Now consider the correlation between G_j and X, and between G_j and Y:

$$\left(\frac{\mathrm{corr}(G_j, X)}{\mathrm{corr}(G_j, Y)}\right)^2 = \frac{\mathrm{cov}^2(G_j, X)\,\mathrm{Var}(Y)}{\mathrm{cov}^2(G_j, Y)\,\mathrm{Var}(X)} = \left(\frac{\alpha_{2j}\beta_{YX} + \alpha_{1j}}{\alpha_{1j}\beta_{XY} + \alpha_{2j}}\right)^2 \frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}. \tag{4.7}$$

Suppose G_j is a valid instrument for X, i.e., $\alpha_{1j} \neq 0$ and $\alpha_{2j} = 0$. Then,

$$\left(\frac{\mathrm{corr}(G_j, X)}{\mathrm{corr}(G_j, Y)}\right)^2 = \frac{1}{\beta_{XY}^2} \cdot \frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)} > 1.$$

Similarly, if G_j is a valid instrument for Y, i.e., $\alpha_{2j} \neq 0$ and $\alpha_{1j} = 0$, then

$$\left(\frac{\mathrm{corr}(G_j, X)}{\mathrm{corr}(G_j, Y)}\right)^2 = \beta_{YX}^2 \cdot \frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)} < 1.$$

Based on these results, we establish the following proposition:

Proposition 1. *For each G_j, $j = 1, \ldots, p$, if $|corr(G_j, X)| < |corr(G_j, Y)|$, then G_j cannot be a valid instrument for X. Conversely, if $|corr(G_j, X)| > |corr(G_j, Y)|$, then G_j cannot be a valid instrument for Y.*

In stage two, prior to applying best subset selection to identify invalid IVs, we incorporate the insights from Proposition 1 to improve both computational efficiency and accuracy. Specifically, for each selected IV G_j, where $j \in \hat{S}_1 = \hat{S}_{X1} \cup \hat{S}_{Y1}$, we compute its empirical correlation with both X and Y. If $|\text{corr}(G_j, X)| < |\text{corr}(G_j, Y)|$, label G_j as invalid for X, if $|\text{corr}(G_j, X)| > |\text{corr}(G_j, Y)|$, label G_j as invalid for Y. Let $\hat{A}_{X1}$ denote the set of pre-identified invalid IVs for X, i.e., $\hat{A}_{X1} = \{j : |\text{corr}(G_j, X)| < |\text{corr}(G_j, Y)|, j \in \hat{S}_{X1}\}$. And let $\hat{A}_{Y1}$ denote the set of pre-identified invalid IVs for Y. Then the remaining IVs are $\hat{S}_{X2} = \hat{S}_{X1} \setminus \hat{A}_{X1}$ for X and $\hat{S}_{Y2} = \hat{S}_{Y1} \setminus \hat{A}_{Y1}$ for Y.

So now we could reformulate Problem 4.3 as:

$$\hat{\alpha}_{2,\hat{S}_{X2}} = \arg\min_{\alpha_{2,\hat{S}_{X2}}} \|\alpha_{2,\hat{S}_{X2}}\|_0 \quad \text{s.t.} \quad \|\widetilde{Y} - \widetilde{\widetilde{G}}_{\hat{S}_{X2}} \alpha_{2,\hat{S}_{X2}}\|_2^2 < \delta, \tag{4.8}$$

where $\widetilde{Y} = M_{\widetilde{G}_{\hat{A}_{X1}}} Y = (I - \widetilde{G}_{\hat{A}_{X1}} (\widetilde{G}'_{\hat{A}_{X1}} \widetilde{G}_{\hat{A}_{X1}})^{-1} \widetilde{G}'_{\hat{A}_{X1}})Y$, $\widetilde{\widetilde{G}}_{\hat{S}_{X2}} = M_{\widetilde{G}_{\hat{A}_{X1}}} \widetilde{G}_{\hat{S}_{X2}}$.

Similarly, reformulate Problem 4.4 as:

$$\hat{\alpha}_{2,\hat{S}_{Y2}} = \arg\min_{\alpha_{2,\hat{S}_{Y2}}} \|\alpha_{2,\hat{S}_{Y2}}\|_0 \quad \text{s.t.} \quad \|\widetilde{X} - \widetilde{\widetilde{G}}_{\hat{S}_{Y2}} \alpha_{2,\hat{S}_{Y2}}\|_2^2 < \delta, \tag{4.9}$$

where $\widetilde{X} = M_{\widetilde{G}_{\hat{A}_{Y1}}} X = (I - \widetilde{G}_{\hat{A}_{Y1}} (\widetilde{G}'_{\hat{A}_{Y1}} \widetilde{G}_{\hat{A}_{Y1}})^{-1} \widetilde{G}'_{\hat{A}_{Y1}})X$, $\widetilde{\widetilde{G}}_{\hat{S}_{Y2}} = M_{\widetilde{G}_{\hat{A}_{Y1}}} \widetilde{G}_{\hat{S}_{Y2}}$.

In summary, based on the criterion established in Proposition 1, we pre-identify a subset of clearly invalid IVs. We then only select invalid IVs within the remaining, ambiguous instruments. This targeted screening step significantly reduces the search space, leading to substantial improvements in computational speed and estimation performance.

We summarize the algorithm in Algorithm 4.1.

4.3 Simulation Studies

In the simulation study, we aim to mimic realistic scenarios as closely as possible. For each replicate, we generate a total of 1000 SNPs, among which 30 are randomly selected to be associated with either X or Y. This implies that 970 SNPs are irrelevant and serve as noise variables. The

Algorithm 4.1 The analytical procedure of BiMR-SPLIT+.

```
Input: $\{X, Y, G\}$, $X, Y \in \mathbb{R}^{N \times 1}$, $G \in \mathbb{R}^{N \times p}$, $p \gg N$.
Output: The causal effect $\hat{\beta}$.
Employ screening methods, such as SIS, to reduce the number of SNPs to a manageable size
  that less than N, typically a hundred.
For $t = 1, \dots, T$
    • Get estimated $\hat{X}$ and $\hat{Y}$ by splitting the sample into two evenly subsets. Denote $\hat{S}_{X1}$ and
      $\hat{S}_{Y1}$ as the selected IV sets, respectively.
    • For each selected SNP $j \in \hat{S}_{X1} \cup \hat{S}_{Y1}$, if $|\text{corr}(G_j, X)| < |\text{corr}(G_j, Y)|$, label $G_j$ as
      invalid for $X$, if $|\text{corr}(G_j, X)| > |\text{corr}(G_j, Y)|$, label $G_j$ as invalid for $Y$; The remaining
      undefined IVs are denoted as $\hat{S}_{X2}$ and $\hat{S}_{Y2}$, respectively.
    • Starting from $k = 0$, set $\alpha_{2,\hat{S}_{X2},k} = \arg\min_{\alpha_{2,\hat{S}_{X2},k}} \|\widetilde{Y} - \widetilde{\widetilde{G}}_{\hat{S}_{X2}} \alpha_{2,\hat{S}_{X2},k}\|_2^2$, then perform the
      overidentification test. If the testing p-value $\leq 0.05$, increment $k$ by 1 (i.e., $k = k + 1$)
      and repeat the process until the testing p-value $p > 0.05$. And similarly for $\alpha_{1,\hat{S}_{X2}}$.
    • Regress Y on $\hat{X}$ while including the selected invalid IVs as covariates to get the estimated
      causal effect $\hat{\beta}_{XY,t}$.
    • Regress X on $\hat{Y}$ while including the selected invalid IVs as covariates to get the estimated
      causal effect $\hat{\beta}_{YX,t}$.
Perform the same sample split procedure $T$ times and choose
      $$\hat{\beta}_{XY} = \arg \min_{\hat{\beta}_{XY,t} \in \{\hat{\beta}_{XY,t} : t = 1, \dots, T\}} \left| \hat{\beta}_{XY,t} - \text{median}(\{\hat{\beta}_{XY,t} : t = 1, \dots, T\}) \right|,$$
      $$\hat{\beta}_{YX} = \arg \min_{\hat{\beta}_{YX,t} \in \{\hat{\beta}_{YX,t} : t = 1, \dots, T\}} \left| \hat{\beta}_{YX,t} - \text{median}(\{\hat{\beta}_{YX,t} : t = 1, \dots, T\}) \right|,$$
  as the final causal effect estimates.
```

phenotypes X and Y are generated according to Model 4.6. The matrix of genotypes G is simulated as independent discrete variables taking values in $\{0, 1, 2\}$, representing the allele count under an additive model, with a fixed MAF of 0.3. The error terms ε_{1i} and ε_{2i} are independently generated across individuals from a bivariate normal distribution with zero mean, unit variances, and a correlation of 0.3, that is,

$$(\varepsilon_{1i}, \varepsilon_{2i})^\top \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right), \quad i = 1, \dots, N.$$

This correlation reflects the presence of unmeasured confounding between X and Y. We consider three different sample sizes in the simulation study, with $N \in \{1000, 2000, 4000\}$.

To reflect different causal structures commonly encountered in practice, we consider the following three settings for the bidirectional causal effects:

- $\beta_{XY} = 0, \beta_{YX} = 0$ (no causal relationship);
- $\beta_{XY} = 0.75, \beta_{YX} = 0$ (unidirectional causality from X to Y);
- $\beta_{XY} = 0.5, \beta_{YX} = 1$ (bidirectional causality).

For the direct effects of SNPs on the phenotypes, we consider two distinct configurations of the vectors α_1 and α_2:

- **Scenario 1:**

$$\alpha_1 = (\underbrace{0.4, \cdots, 0.4}_{7}, \underbrace{0, \cdots, 0}_{7}, \underbrace{0.4, \cdots, 0.4}_{4}, \underbrace{0.3, \cdots, 0.3}_{4}, \underbrace{0.2, \cdots, 0.2}_{4}, 0.1, 0.1, 0.4, 0.4),$$

$$\alpha_2 = (\underbrace{0, \cdots, 0}_{7}, \underbrace{0.4, \cdots, 0.4}_{7}, \underbrace{0.4, \cdots, 0.4}_{4}, \underbrace{0.2, \cdots, 0.2}_{4}, \underbrace{0.3, \cdots, 0.3}_{4}, 0.4, 0.4, 0.1, 0.1).$$

- **Scenario 2:**

$$\alpha_1 = (\underbrace{0.4, \cdots, 0.4}_{7}, \underbrace{0, \cdots, 0}_{7}, \underbrace{0.4, \cdots, 0.4}_{4}, \underbrace{0.3, \cdots, 0.3}_{4}, \underbrace{0.2, \cdots, 0.2}_{4}, \underbrace{0.1, \cdots, 0.1}_{4}),$$

$$\alpha_2 = (\underbrace{0, \cdots, 0}_{7}, \underbrace{0.4, \cdots, 0.4}_{7}, \underbrace{0.4, \cdots, 0.4}_{4}, \underbrace{0.2, \cdots, 0.2}_{4}, \underbrace{0.3, \cdots, 0.3}_{4}, \underbrace{0.1, \cdots, 0.1}_{4}).$$

Scenario 1 satisfies the plurality rule when considering unidirectional MR under no bidirectional causality. That is, the largest group of selected IVs that generate the same causal effects are the valid ones.

Scenario 2 represents a more challenging setting in which the plurality rule is violated regardless of whether bidirectional causality is present. For instance, under the no causal relationship setting ($\beta_{XY} = \beta_{YX} = 0$), when estimating β_{XY}, the first 7 SNPs are valid instruments with $\alpha_1 = 0.4$ and $\alpha_2 = 0$, leading to an estimated causal effect of $\hat{\beta}_{XY} = 0$. However, there are 8 other SNPs that contribute to an estimated $\hat{\beta}_{XY} = 1$, including 4 SNPs with $\alpha_1 = \alpha_2 = 0.4$ and 4 with $\alpha_1 = \alpha_2 = 0.1$.

As a result, the invalid instruments dominate numerically, and the plurality of instruments support an incorrect causal direction, thereby violating the plurality rule. It is designed to evaluate the robustness of the BiMR-SPLIT+.

We compare the performance of BiMR-SPLIT+ against the following benchmark methods to demonstrate its superiority:

- **Oracle TSLS**, which assumes complete knowledge of the valid and invalid IVs;
- **CIIV**, a consistent IV selection and estimation method;
- **MR-Egger**, a widely used method for addressing directional pleiotropy in Mendelian randomization.

4.3.1 Simulation results of scenario 1

Table A.7 in the Appendix presents the simulation results for Scenario 1, where both $\beta_{XY} = 0$ and $\beta_{YX} = 0$. This setting reflects a null causal relationship in both directions. Among the non-oracle methods, BiMR-SPLIT+ consistently yields bias estimates closest to those of Oracle TSLS, especially as sample size increases. For example, the bias for BiMR-SPLIT+ reduces from 0.0495 at $N = 1000$ to 0.0030 at $N = 4000$ in the $X \rightarrow Y$ direction, and from 0.0498 to –0.0028 in the $Y \rightarrow X$ direction.

BiMR-SPLIT+ also achieves lower RMSE compared to MR-Egger and CIIV across all settings, demonstrating greater estimation precision. While its coverage probability (CP) is initially below the nominal 95%, it improves with larger sample size, from 0.68 to 0.87 in the $X \rightarrow Y$ direction and from 0.65 to 0.94 in the $Y \rightarrow X$ direction, indicating that the method becomes increasingly reliable as data size grows. In addition, we report the false positive rate (FPR) and false negative rate (FNR), which respectively measure the proportion of invalid IVs that are incorrectly retained and valid IVs that are incorrectly excluded. These two results of BiMR-SPLIT+ both are decreasing as the sample size increasing, which also shows the validity of identifying invalid IVs for this method.

In contrast, MR-Egger exhibits high coverage but at the expense of large RMSE and much wider confidence intervals (e.g., width up to 1.28), and tends to produce more biased estimates, especially under small sample sizes. Although the plurality rule is not violated in this scenario, CIIV still performs poorly, with extremely high bias, RMSE approaching or exceeding 0.9, and very low coverage (as low as 6%–44%). Its performance improves as the sample size increases: bias gradually decreases, and FPR/FNR indicate more accurate instrument classification. This suggests that CIIV requires either a sufficiently large sample size or strong instruments to reliably distinguish between different groups of IVs. When the group separation is weak, the method struggles to identify the correct set of valid instruments, leading to poor estimation performance.

When considering the setting $\beta_{XY} = 0.75$ and $\beta_{YX} = 0$, the advantages of BiMR-SPLIT+ become even more evident, see Table 4.1. In both directions ($X \longrightarrow Y$ and $Y \longrightarrow X$), BiMR-SPLIT+ consistently ranks as the second-best method after the Oracle TSLS, and its performance becomes increasingly comparable to the oracle estimator as the sample size grows. Notably, the FPR and FNR for classifying invalid and valid IVs approach zero with larger sample sizes, clearly demonstrating the superiority of BiMR-SPLIT+ over CIIV, whose classification ability nearly fails. Additionally, the estimation bias of BiMR-SPLIT+ diminishes, and the coverage probability (CP) converges steadily toward the nominal level of 95%.

For comparison, MR-Egger suffers from substantial bias, and even with its overly wide confidence intervals, it still fails to attain acceptable coverage in the $Y \longrightarrow X$ direction. CIIV also remains suboptimal, with both FPR and FNR failing to reach ideal levels, indicating that its underlying assumptions for effectively identifying valid instruments are more stringent than those required by BiMR-SPLIT+, and may be difficult to satisfy in practical applications.

Table 4.2 are the results when $\beta_{XY} = 0.5$ and $\beta_{YX} = 1$. Under this setting, BiMR-SPLIT+ exhibits similar performance as in previous scenarios. Although in the $X \longrightarrow Y$ direction with $N = 1000$, the coverage probability (CP) does not reach the nominal 95% and its FPR is 0.03, both metrics improve rapidly with larger sample size. When $N = 2000$, the FPR drops to 0.00 and the CP rises to 100%, matching the performance of the Oracle TSLS. This result highlights two aspects: first, even a small fraction of misclassified invalid IVs can lead to considerable estimation bias; second, BiMR-SPLIT+ demonstrates high efficiency in identifying valid and invalid instruments.

Table 4.1 Simulation results of scenario 1 when $\beta_{XY} = 0.75, \beta_{YX} = 0$.

Settings	N	Method	Bias	Est.sd	RMSE	CI Width	CP	FPR	FNR
$\beta_{XY} = 0.75$	1000	Oracle TSLS	0.0019	0.0219	0.0220	0.1295	0.99	0.00	0.00
		BiMR-SPLIT+	0.0332	0.0466	0.0571	0.1690	0.84	0.13	0.01
		MR-Egger	-0.2152	0.2155	0.3044	1.2879	0.98	NA	NA
		CIIV	0.9118	0.3744	0.9856	0.1960	0.11	0.63	0.88
	2000	Oracle TSLS	0.0000	0.0156	0.0156	0.0913	1.00	0.00	0.00
		BiMR-SPLIT+	0.0107	0.0367	0.0382	0.1209	0.88	0.06	0.01
		MR-Egger	-0.1938	0.1574	0.2496	1.1001	0.99	NA	NA
		CIIV	0.7027	0.7205	1.0059	0.1797	0.44	0.28	0.52
	4000	Oracle TSLS	0.0002	0.0119	0.0119	0.0642	0.99	0.00	0.00
		BiMR-SPLIT+	-0.0004	0.0236	0.0236	0.0884	0.94	0.01	0.00
		MR-Egger	-0.1766	0.1360	0.2229	1.0024	1.00	NA	NA
		CIIV	0.5569	0.7739	0.9529	0.1293	0.62	0.17	0.35
$\beta_{YX} = 0$	1000	Oracle TSLS	0.0021	0.0233	0.0234	0.0914	0.95	0.00	0.00
		BiMR-SPLIT+	-0.0047	0.0296	0.0299	0.0881	0.86	0.02	0.01
		MR-Egger	0.3844	0.1214	0.4031	0.7808	0.50	NA	NA
		CIIV	0.3990	0.2800	0.4873	0.0737	0.17	0.30	0.80
	2000	Oracle TSLS	0.0003	0.0162	0.0162	0.0644	0.96	0.00	0.00
		BiMR-SPLIT+	-0.0079	0.0177	0.0194	0.0642	0.91	0.00	0.01
		MR-Egger	0.3450	0.0982	0.3587	0.7902	0.67	NA	NA
		CIIV	0.4628	0.3577	0.5847	0.0575	0.14	0.28	0.84
	4000	Oracle TSLS	0.0002	0.0114	0.0114	0.0453	0.95	0.00	0.00
		BiMR-SPLIT+	-0.0039	0.0115	0.0121	0.0453	0.94	0.00	0.00
		MR-Egger	0.3406	0.0693	0.3475	0.8062	0.74	NA	NA
		CIIV	0.5094	0.5394	0.7415	0.0510	0.34	0.19	0.62

CIIV and MR-Egger show performance patterns consistent with earlier settings. For CIIV, insufficient accuracy in identifying invalid IVs leads to substantial bias. For MR-Egger, the excessively wide confidence intervals result in low estimation precision, despite achieving relatively high coverage in some cases.

4.3.2 Simulation results of scenario 2

Table A.8 in the Appendix presents the simulation results for Scenario 2 under the no causal relationship setting, where $\beta_{XY} = \beta_{YX} = 0$. Across all sample sizes ($N = 1000, 2000, 4000$), BiMR-SPLIT+ achieves consistently low bias and RMSE. For instance, when $N = 1000$, BiMR-SPLIT+ has a bias of 0.0328 and RMSE of 0.1364, substantially lower than MR-Egger (bias = –0.1046, RMSE = 0.2440) and CIIV (bias = 0.8653, RMSE = 0.9387).

BiMR-SPLIT+ also maintains coverage probabilities around 90%, with moderate confidence

Table 4.2 Simulation results of scenario 1 when $\beta_{XY} = 0.5, \beta_{YX} = 1$.

Settings	N	Method	Bias	Est.sd	RMSE	CI Width	CP	FPR	FNR
$\beta_{XY} = 0.5$	1000	Oracle TSLS	0.0023	0.0108	0.0111	0.1131	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0342	0.0208	0.0400	0.1080	0.86	0.03	0.00
		MR-Egger	0.2290	0.0468	0.2337	0.3094	0.06	NA	NA
		CIIV	0.3200	0.1149	0.3400	0.0302	0.04	0.34	0.95
	2000	Oracle TSLS	0.0007	0.0078	0.0078	0.0801	1.00	0.00	0.00
		BiMR-SPLIT+	0.0021	0.0078	0.0081	0.0797	1.00	0.00	0.00
		MR-Egger	0.2317	0.0383	0.2348	0.3122	0.04	NA	NA
		CIIV	0.3003	0.1399	0.3312	0.0231	0.12	0.29	0.87
	4000	Oracle TSLS	0.0004	0.0060	0.0060	0.0563	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0124	0.0065	0.0140	0.0565	0.99	0.00	0.00
		MR-Egger	0.2275	0.0245	0.2288	0.3189	0.01	NA	NA
		CIIV	0.2721	0.2067	0.3416	0.0197	0.30	0.21	0.68
$\beta_{YX} = 1$	1000	Oracle TSLS	0.0007	0.0117	0.0117	0.1471	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0403	0.0228	0.0463	0.1409	0.94	0.05	0.00
		MR-Egger	0.0781	0.0732	0.1070	0.4894	0.98	NA	NA
		CIIV	0.2045	0.1522	0.2548	0.0454	0.22	0.31	0.75
	2000	Oracle TSLS	0.0000	0.0081	0.0081	0.1037	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0230	0.0116	0.0258	0.1043	1.00	0.01	0.00
		MR-Egger	0.0261	0.0564	0.0621	0.4385	1.00	NA	NA
		CIIV	0.2709	0.2674	0.3805	0.0406	0.16	0.28	0.82
	4000	Oracle TSLS	0.0000	0.0057	0.0057	0.0730	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0123	0.0061	0.0137	0.0736	1.00	0.00	0.00
		MR-Egger	0.0145	0.0470	0.0492	0.4406	1.00	NA	NA
		CIIV	0.2953	0.3888	0.4879	0.0370	0.41	0.17	0.55

interval widths (e.g., 0.154 at $N = 1000$ for $\hat{\beta}_{XY}$). In contrast, MR-Egger, although achieving perfect coverage (e.g., 0.99–1.00), does so at the cost of much wider confidence intervals (e.g., from 1.04 to 1.30), indicating low estimation efficiency.

CIIV performs particularly poorly in this null scenario, with severe bias and extremely low coverage probabilities. The FNRs for CIIV approach 0.86–0.94 across all settings, indicating that the method fails to preserve the majority of valid instruments. This failure is primarily due to the violation of the plurality rule under Scenario 2, where the largest group of IVs no longer represents the valid IVs.

These results indicate that BiMR-SPLIT+ achieves a favorable trade-off between bias, confidence interval efficiency, and coverage, offering reliable Type I error control without being overly conservative like MR-Egger or unstable like CIIV. Its low false discovery rates (e.g., FPR =

0.032–0.10 and FNR = 0.029–0.038) further highlight its robustness in invalid IV selection.

Table 4.3 reports the simulation results for Scenario 2 under the setting $\beta_{XY} = 0.75$ and $\beta_{YX} = 0$, representing the causal effect only from X to Y.

For the direction from X to Y, across all sample sizes ($N = 1000, 2000, 4000$), BiMR-SPLIT+ yields highly accurate estimates of β_{XY}, with biases close to zero and consistently low RMSE values. While MR-Egger achieves nominal coverage near 1.0, it suffers from significantly inflated RMSE and wide confidence intervals. CIIV, on the other hand, completely fails in this setting due to violation of the plurality rule, resulting in substantial bias, RMSE exceeding 0.93, and coverage below 12%.

In the reverse direction from Y to X, where no causal effect exists, BiMR-SPLIT+ maintains low bias and coverage rates close to 90%. In contrast, MR-Egger fails to estimate β_{YX} accurately, with biases exceeding 0.34, high RMSE values, and coverage probabilities near 60%, indicating serious false positives. CIIV again performs poorly, with false positive rates (FPR) exceeding 30–40% and false negative rates (FNR) close to 1.0, reflecting its inability to select valid instruments under this challenging structure.

Table 4.4 reports the simulation results for Scenario 2 under the bidirectional causal setting, where $\beta_{XY} = 0.5$ and $\beta_{YX} = 1.0$. This is a particularly challenging case, as invalid IVs are present for both directions and the plurality rule is violated.

For the direction from $X \to Y$, BiMR-SPLIT+ achieves moderate to good estimation performance. Although the bias is somewhat inflated at $N = 1000$ (–0.0724), it improves with increasing sample size, reaching –0.0117 at $N = 4000$. RMSE decreases steadily from 0.0883 to 0.0133, and coverage increases from 67% to 100%. Similar to earlier scenarios, although MR-Egger maintains high coverage across all settings, it suffers from excessively wide confidence intervals, limiting its estimation efficiency.

For the reverse direction $Y \to X$, where the true causal effect is stronger ($\beta_{YX} = 1.0$), BiMR-SPLIT+ remains accurate and stable. The bias decreases from –0.0557 at $N = 1000$ to –0.0126 at $N = 4000$, with corresponding reductions in RMSE. In contrast, MR-Egger exhibits substantial

Table 4.3 Simulation results of scenario 2 when $\beta_{XY} = 0.75, \beta_{YX} = 0$.

Settings	N	Method	Bias	Est.sd	RMSE	CI Width	CP	FPR	FNR
$\beta_{XY} = 0.75$	1000	Oracle TSLS	0.0019	0.0219	0.0220	0.1295	0.99	0.00	0.00
		BiMR-SPLIT+	-0.0050	0.0443	0.0445	0.1808	0.96	0.07	0.00
		MR-Egger	-0.1006	0.2213	0.2429	1.3059	0.99	NA	NA
		CIIV	0.8653	0.3642	0.9387	0.1865	0.12	0.69	0.86
	2000	Oracle TSLS	0.0000	0.0156	0.0156	0.0913	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0083	0.0289	0.0300	0.1275	0.97	0.04	0.01
		MR-Egger	-0.0143	0.1435	0.1441	1.1317	1.00	NA	NA
		CIIV	0.8815	0.3276	0.9403	0.1345	0.11	0.60	0.89
	4000	Oracle TSLS	0.0002	0.0119	0.0119	0.0642	0.99	0.00	0.00
		BiMR-SPLIT+	-0.0035	0.0203	0.0206	0.0895	0.97	0.01	0.00
		MR-Egger	0.0271	0.1112	0.1144	1.0386	1.00	NA	NA
		CIIV	0.9272	0.2663	0.9646	0.1028	0.06	0.52	0.93
$\beta_{YX} = 0$	1000	Oracle TSLS	0.0021	0.0233	0.0234	0.0914	0.95	0.00	0.00
		BiMR-SPLIT+	-0.0071	0.0711	0.0714	0.0909	0.88	0.02	0.03
		MR-Egger	0.3500	0.1157	0.3686	0.7787	0.62	NA	NA
		CIIV	0.5651	0.2502	0.6179	0.0580	0.07	0.39	0.92
	2000	Oracle TSLS	0.0003	0.0162	0.0162	0.0644	0.96	0.00	0.00
		BiMR-SPLIT+	-0.0030	0.0589	0.0589	0.0643	0.91	0.00	0.02
		MR-Egger	0.3383	0.0830	0.3483	0.7305	0.62	NA	NA
		CIIV	0.5607	0.1757	0.5875	0.0373	0.05	0.35	0.95
	4000	Oracle TSLS	0.0002	0.0114	0.0114	0.0453	0.95	0.00	0.00
		BiMR-SPLIT+	0.0116	0.0886	0.0892	0.0452	0.92	0.01	0.03
		MR-Egger	0.3451	0.0560	0.3496	0.7232	0.56	NA	NA
		CIIV	0.5776	0.0743	0.5824	0.0236	0.00	0.35	1.00

bias and poor coverage (as low as 2%), indicating that it may fail to provide reliable estimates in this direction. CIIV again performs poorly, with severe bias, low coverage, and high false negative rates, confirming its ineffectiveness under this setting.

In summary, BiMR-SPLIT+ consistently demonstrates strong performance in terms of estimation accuracy, robustness, and instrument selection. Especially in the most challenging bidirectional causal scenario, where causal effects exist for both directions and the plurality rule is violated, BiMR-SPLIT+ remains the only method that achieves stable and accurate estimation in both directions. As sample size increases, its bias and RMSE steadily decrease and coverage improves to approach or reach the nominal level. Taken together, these results highlight the robustness and adaptability of BiMR-SPLIT+ across a wide range of causal structures and IV validity conditions, making it a practical and reliable tool for bidirectional MR analysis.

Table 4.4 Simulation results of scenario 2 when $\beta_{XY} = 0.5, \beta_{YX} = 1$.

Settings	N	Method	Bias	Est.sd	RMSE	CI Width	CP	FPR	FNR
$\beta_{XY} = 0.5$	1000	Oracle TSLS	0.0016	0.0109	0.0110	0.1468	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0724	0.0422	0.0838	0.1560	0.67	0.02	0.01
		MR-Egger	0.0523	0.0699	0.0873	0.4758	1.00	NA	NA
		CIIV	0.2996	0.1377	0.3296	0.0380	0.08	0.41	0.89
	2000	Oracle TSLS	0.0004	0.0078	0.0078	0.1037	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0255	0.0093	0.0271	0.1058	1.00	0.00	0.00
		MR-Egger	0.0457	0.0461	0.0649	0.4119	1.00	NA	NA
		CIIV	0.3352	0.1620	0.3722	0.0279	0.06	0.35	0.93
	4000	Oracle TSLS	0.0003	0.0060	0.0060	0.0729	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0117	0.0064	0.0133	0.0736	1.00	0.00	0.00
		MR-Egger	0.0461	0.0373	0.0592	0.4006	1.00	NA	NA
		CIIV	0.3303	0.0347	0.3321	0.0158	0.01	0.34	0.99
$\beta_{YX} = 1$	1000	Oracle TSLS	0.0014	0.0116	0.0117	0.1135	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0557	0.0217	0.0598	0.1166	0.57	0.01	0.01
		MR-Egger	0.2533	0.0459	0.2574	0.3111	0.02	NA	NA
		CIIV	0.2482	0.0989	0.2672	0.0223	0.06	0.39	0.93
	2000	Oracle TSLS	0.0003	0.0081	0.0081	0.0801	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0263	0.0104	0.0283	0.0811	0.94	0.00	0.00
		MR-Egger	0.2477	0.0331	0.2499	0.2926	0.00	NA	NA
		CIIV	0.2446	0.0712	0.2547	0.0145	0.05	0.35	0.95
	4000	Oracle TSLS	0.0002	0.0057	0.0057	0.0563	1.00	0.00	0.00
		BiMR-SPLIT+	-0.0126	0.0063	0.0141	0.0566	1.00	0.00	0.00
		MR-Egger	0.2501	0.0200	0.2509	0.2882	0.00	NA	NA
		CIIV	0.2512	0.0232	0.2522	0.0090	0.00	0.35	1.00

4.4 Application: Causal Pathway Between Gene Expression and Trait

In this section, we demonstrate the practical utility of our method by applying it to a real-world biological dataset. For now, large-scale individual-level datasets are currently limited in availability, so we illustrate our approach using a dataset comprising 200 male Drosophila melanogaster samples as a representative case study. The primary phenotype of interest is phototaxis, which was measured at two time points: day 4 and day 28 of age. In parallel, gene expression profiles were obtained at both 1 week and 4 weeks of age. Our goal is to accurately classify gene–trait relationships according to their causal direction: distinguishing which gene expressions act as causal drivers of phototactic behavior and which represent reactive responses. By separately analyzing data from young and aged flies, we aim to uncover age-specific biological mechanisms underlying phototactic regulation and demonstrate the method's capacity to resolve directionality in observational transcriptomicphenotypic associations.

For the gene dataset, We first performed standard quality control procedures on genotype data from 200 Drosophila melanogaster lines. Variants with a high missing genotype rate (> 10%) were removed, and only common variants with a MAF ≥ 0.05 were retained. We then conducted LD pruning using a sliding window approach with an r^2 threshold of 0.64, resulting in a final dataset of 931,732 approximately independent SNPs for downstream analysis.

After merging the available gene expression datasets, we obtained expression profiles for 12,510 genes across 180 one-week-old (young) Drosophila samples and 12,361 genes across 176 four-week-old (aged) samples. To focus on genes most relevant to the behavioral phenotype of interest, we performed a marginal association analysis between gene expression and phototaxis. Given the relatively small sample sizes, we adopted a liberal screening threshold of marginal p-value < 0.01 to retain potentially informative features. This resulted in the selection of 71 genes in the young group and 64 genes in the aged group for downstream analysis, with only one gene expression 'FBgn0035932' shared between the two groups.

Next, we applied the BiMR-SPLIT+ method to perform bidirectional MR analyses between each selected gene expression variable and the phototaxis phenotype in both age groups. Given the limited sample sizes, we performed 60 random splits when applying BiMR-SPLIT+ to enhance the stability and reliability of the results.

4.4.1 Young Group Results

Table 4.5 presents the significant gene expression identified by applying BiMR-SPLIT+ to the young group (i.e., one-week-old Drosophila).

The first columns list the genes identified as causal drivers of phototaxis. Among them, the most significant finding is FBgn0003733, which exhibits a negative causal effect on phototaxis. FBgn0003733 corresponds to the torso (tor) gene in Drosophila melanogaster, which encodes a receptor protein-tyrosine kinase known for its role in embryonic patterning and hormonal regulation during metamorphosis. Notably, during the larval stage, Torso acts as the receptor for prothoracicotropic hormone (PTTH), which is a key neuroendocrine signal that initiates a cascade controlling

Table 4.5 Identified Significant Gene Expressions in Young Flies.

Gene Expression	Estimate	Std. Error	p-value	Lower bound	Upper bond	Causal Mechanism
FBgn0003733	-3.35	1.23	0.0069	-5.75	-0.95	Causal
FBgn0031468	1.17	0.45	0.0100	0.29	2.04	Causal
FBgn0085273	0.95	0.37	0.0103	0.23	1.67	Causal
FBgn0039066	10.12	4.42	0.0231	1.47	18.78	Causal
FBgn0043364	3.59	1.58	0.0244	0.49	6.69	Causal
FBgn0267819	1.50	0.66	0.0257	0.19	2.80	Causal
FBgn0000279	0.11	0.04	0.0133	0.02	0.19	Reactive
FBgn0033781	-0.01	0.01	0.0236	-0.03	0.00	Reactive
FBgn0021765	-0.01	0.00	0.0314	-0.02	0.00	Reactive
FBgn0026576	-0.02	0.01	0.0320	-0.04	0.00	Reactive

light avoidance behavior (i.e., negative phototaxis), as demonstrated by Yamanaka et al. (2013). Mechanistically, PTTH binds to the Torso receptor and activates downstream signaling pathways that modulate the function of two major light-sensing systems: the Bolwig's organ and class IV dendritic arborization neurons. These sensory neurons detect ambient light and are inhibited or modulated by Torso signaling, ultimately promoting larval movement toward darker environments as they prepare for pupation. This neuroendocrine-driven behavioral adaptation enhances survival by ensuring that larvae enter the pupal stage in protective, low-light environments.

Furthermore, five other gene expressions were identified as having significant causal effects in promoting phototactic behavior. FBgn0031468, corresponding to CG2975 (Müller et al., 2005), encodes a β-1,3-galactosyltransferase involved in protein O-glycosylation, which is a critical post-translational modification affecting membrane and synaptic protein function. Although it has not been directly linked to phototaxis previously, altered glycosylation in light-sensing neurons (Katoh and Tiemeyer, 2013) may affect receptor stability, localization, or signaling efficiency, ultimately modulating behavioral response to light. Its enriched expression in early development and adult males (Brown et al., 2014) further supports its potential contribution to phototactic variation in this age group. FBgn0039066 encodes EloA, the active subunit of the Elongin complex, which facilitates transcriptional elongation by RNA polymerase II (Gerber et al., 2004). EloA is highly expressed in early embryos (Brown et al., 2014) and is localized to central brain structures, suggesting its involvement in neural gene regulation. Its upregulation may accelerate the transcription of genes critical to phototransduction, synaptic plasticity, or sensory processing, thereby enhancing responsiveness to light stimuli and supporting phototactic behavior. FBgn0043364, also known as cabut (cbt), encodes a zinc-finger transcription factor implicated in BMP signaling and sensory organ development (Mukherjee et al., 2021). Given its high expression during late embryogenesis and established roles in neurogenesis, cbt likely supports the differentiation and connectivity of photoreceptive circuits (Abdelilah-Seyfried et al., 2000). Its causal effect on phototaxis may arise through developmental programming of light-sensitive neural systems. For the remaining two genes, FBgn0085273 and FBgn0267819, no well-characterized links to phototactic behavior or neuro development have been established. However, their reproducible causal relationship with phototaxis in this dataset suggests that they may represent novel regulatory components or indirect modulators of light-responsive behavior. Further investigation into their function and expression dynamics is warranted.

In addition, four gene expressions were identified as reactive responses. First, phototactic behavior was found to positively regulate the expression of Cecropin C (CecC, FBgn0000279), an antimicrobial peptide gene involved in the innate immune response (TRYSELIUS et al., 1992; Gordon et al., 2008; Carboni et al., 2022; Verleyen et al., 2006). This upregulation may reflect an anticipatory immune response triggered by increased environmental exploration. Young flies exhibiting higher phototaxis are likely more active and more exposed to microbial threats in external environments. As a result, the innate immune system may be primed through behavior-linked signals to express effector genes such as CecC, which encodes a peptide active against Gram-negative bacteria. Moreover, increased behavioral engagement may activate neuroendocrine signaling cascades (e.g., Imd, Toll), which intersect with immune transcriptional networks (Davies et al., 2012).

In contrast, there exist mild suppression of several mitochondrial and cellular maintenance-related genes, FBgn0033781 (CG13319), FBgn0021765 (CG7113, scully), and FBgn0026576 (Pisd), in response to increased phototactic behavior. These genes are functionally distinct but share involvement in core biological processes such as proteasome assembly (CG13319), mitochondrial tRNA processing and steroid metabolism (scu) (Torroja et al., 1998), and phospholipid biosynthesis in the mitochondrial membrane (Pisd) (Zhao and Wang, 2020). This consistent downregulation may reflect a short-term physiological prioritization of sensory and motor functions over background maintenance tasks. Phototaxis is a behavior that requires sustained visual attention and locomotion, which may transiently shift transcriptional activity away from mitochondrial biogenesis and metabolic housekeeping. Such reallocation of resources in young adults likely represents a flexible and reversible trade-off, allowing flies to adapt their molecular programs to immediate behavioral demands. Additionally, enhanced sensory-driven activity could generate mild neural or metabolic stress signals, especially in energy-intensive tissues like the head or thoracic muscles, resulting in temporary downregulation of mitochondrial and quality control genes.

4.4.2 Aged Group Results

In the aged Drosophila, RalGPS (FBgn0034158), PNUTS (FBgn0053526), and CG33673 (FBgn0053673) was found to associated with enhanced phototactic behavior, see Table 4.6.

Table 4.6 Identified Significant Gene Expressions in Aged Flies.

Gene Expression	Estimate	Std. Error	p-value	Lower bound	Upper bond	Causal Mechanism
FBgn0034158	2.42	0.80	0.0029	0.85	3.99	Causal
FBgn0035317	-2.13	0.85	0.0131	-3.80	-0.47	Causal
FBgn0039674	-4.13	1.75	0.0194	-7.57	-0.70	Causal
FBgn0053526	3.84	1.80	0.0347	0.31	7.37	Causal
FBgn0039640	-1.95	0.92	0.0370	-3.76	-0.13	Causal
FBgn0053673	4.90	2.38	0.0410	0.24	9.55	Causal
FBgn0032883	-2.94	1.45	0.0442	-5.78	-0.10	Causal
FBgn0266967	-4.80	2.37	0.0443	-9.45	-0.16	Causal
FBgn0028978	0.02	0.01	0.0198	0.00	0.03	Reactive
FBgn0085227	-0.03	0.01	0.0221	-0.06	0.00	Reactive
FBgn0266819	0.04	0.02	0.0247	0.01	0.08	Reactive
FBgn0083963	0.00	0.00	0.0256	0.00	0.00	Reactive
FBgn0035932	-0.10	0.04	0.0261	-0.19	-0.01	Reactive
FBgn0261058	0.01	0.01	0.0285	0.00	0.03	Reactive
FBgn0004865	0.01	0.00	0.0343	0.00	0.02	Reactive

RalGPS encodes a guanyl-nucleotide exchange factor that regulates Ras/Ral GTPase signaling and epidermal growth factor receptor (EGFR) pathways (Nászai et al., 2021). In aged flies, increased RalGPS activity may boost synaptic plasticity or neural excitability through ERK pathway activation (Ferro and Trabalzini, 2010; Impey et al., 1999). These effects could reinforce visual-motor coupling, enabling stronger behavioral responses to light stimuli. PNUTS (Phosphatase 1 Nuclear Targeting Subunit) is involved in gene expression and developmental regulation (Ciurciu et al., 2013). Its high expression in older flies may help stabilize transcriptional networks needed for maintaining synaptic integrity or sustaining locomotor readiness, counteracting age-related decline in neuromotor coordination. CG33673 is predicted to encode a calcium channel component (Project, 2011), possibly localized to Golgi or plasma membranes. Enhanced calcium signaling in the aged brain can elevate neuronal firing rates and enhance sensory responsiveness (Berridge, 1998). In the context of phototaxis, calcium influx may facilitate visual circuit reactivity or downstream motor output (Brini et al., 2014). Together, these genes may act through distinct but converging pathways to support sensory fidelity and behavioral responsiveness in the aging nervous system.

Additionally, several genes, Oseg2 (FBgn0035317), CG1907 (FBgn0039674), superdeath (FBgn0039640), Rhau (FBgn0032883), and the CR45418 (FBgn0266967), were found to suppress phototactic behavior. Oseg2 is crucial for intraflagellar transport and the maintenance of sensory cilium structure (Avidor-Reiss et al., 2004). CG1907 is predicted to encode a mitochondrial dicarboxylate transporter involved in the malate-aspartate shuttle (Gene Ontology Curators, 02). RHAU helicase (Rhau) encodes a protein responsible for G-quadruplex DNA unwinding (You et al., 2017; Lattmann et al., 2010). To date, however, there are no published studies directly reporting inhibitory effects of these genes on phototactic behavior. Our results may thus represent novel findings, suggesting previously unrecognized roles for these genes in the regulation of phototaxis in aged Drosophila. Additionally, the precise biological functions of superdeath and CR45418 in the context of phototaxis remain unknown, as their roles in specific biological processes have yet to be characterized.

Table 4.6 also shows that, increased phototactic activity was found to mildly upregulate the expression of several genes involved in neural function, cellular signaling, and reproductive regulation: FBgn0028978 (tribbles), FBgn0083963 (Neuroligin 3), FBgn0261058 (Seminal fluid protein 38D), and FBgn0004865 (Ecdysone-induced protein 78C).

Tribbles (trbl) encodes a protein kinase inhibitor known to regulate MAP kinase signaling and insulin-like signaling pathways (Das et al., 2014). Enhanced expression of trbl in response to heightened phototactic behavior may reflect increased demands on neural signaling pathways, potentially acting as a feedback mechanism to prevent excessive activation of neuronal signaling cascades and maintain neural homeostasis. Neuroligin 3 (Nlg3) encodes a synaptic adhesion molecule critical for synapse formation, stabilization, and neural transmission (Xing et al., 2014). Elevated phototactic behavior, an activity that requires robust neuronal communication and synaptic plasticity, may drive increased Nlg3 expression to support synaptic strengthening and maintain effective neurotransmission during heightened sensory processing. Seminal fluid protein 38D (Sfp38D) is primarily known for its role in reproductive biology (Findlay et al., 2009). Its upregulation, however, could indicate broader physiological adaptations that link sensory or behavioral activity with reproductive function, possibly mediated via neuroendocrine signals triggered by increased phototactic activity. Ecdysone-induced protein 78C (Eip78C) is predicted to encode a DNA-binding transcription factor involved in regulating gene expression in response to hormonal cues (ecdysone signaling) (Members, 04). Mildly increased expression of Eip78C could reflect phototactic-induced neuroendocrine activation, integrating environmental cues (e.g., increased light exposure) with transcriptional changes necessary for physiological adaptation, stress response, or metabolic adjustments in older flies.

For the other two genes found to be mildly inhibited by phototaxis (FBgn0085227 and FBgn0035932), the underlying biological mechanisms remain unclear. However, our findings may provide a new perspective for future research.

4.5 Discussion

In this study, we have successfully extended the MR-SPLIT+ framework to bidirectional causal inference by developing the BiMR-SPLIT+ method. This new algorithm is specifically designed to address the challenge of invalid instrumental variables (IVs) that arise when the plurality rule fails in the presence of bidirectional causality. At the same time, BiMR-SPLIT+ further improves computational efficiency compared to the original MR-SPLIT+ approach.

Through comprehensive simulation studies, we focused on two particularly challenging scenarios and compared the performance of BiMR-SPLIT+ to that of oracle TSLS, CIIV, and MR-Egger methods. In both settings, BiMR-SPLIT+ consistently provided robust and reliable estimates, exhibiting strong adaptability to complex real-world conditions. Importantly, it produced the lowest bias among all methods except the oracle, while maintaining high coverage probabilities for confidence intervals.

In our empirical application, we successfully applied BiMR-SPLIT+ to a Drosophila dataset with approximately 180 available samples. The method identified gene expressions with causal effects on phototaxis in both young and aged fly cohorts, with many findings corroborated by existing biological literature, thus further validating our approach. As larger-scale individual-level datasets in humans become available, BiMR-SPLIT+ is well-positioned to elucidate the causal mechanisms underlying complex diseases, thereby facilitating the identification of true causal drivers for targeted therapeutic development.

In summary, BiMR-SPLIT+ represents a valuable and generalizable tool for robust bidirectional causal inference in both experimental and observational genomics studies

Looking forward, our proposed framework for bidirectional causality can be naturally extended to the construction of causal networks, offering promising opportunities for elucidating more complex causal mechanisms in future studies. Moreover, owing to the inherent flexibility of this framework, it can also be readily adapted to accommodate nonlinear causal relationships. Such extensions have the potential to uncover more intricate causal structures and yield more accurate causal effect estimates in increasingly complex settings. In addition, there remains room for improvement in the construction of confidence intervals within this framework, which will require further theoretical development.

CHAPTER 5

CONCLUSION AND DISCUSSION

5.1 Summary of Main Contributions

This dissertation makes several key contributions to the methodology and application of Mendelian randomization for causal inference in genomics.

First, we developed the MR-SPLIT method within the 2SLS framework to address two major challenges in one-sample MR analysis: instrument selection bias and the weak instrument problem. MR-SPLIT employs adaptive random sample splitting, using one half of the data for IV selection and the other for causal estimation, thereby avoiding the "winner's curse" from reusing the same sample. We further enhanced robustness through multiple sample splitting and aggregation of weak IVs into composite instruments. Extensive theoretical evaluation and simulation studies show that MR-SPLIT outperforms traditional methods such as 2SLS and LIML, as well as the cross-fitting MR (CFMR) approach, in both bias reduction and statistical power. Empirical analysis with the CRIC dataset further demonstrated its practical utility in establishing the causal role of kidney function on aTRH. In addition, we explored LASSO and de-biased LASSO methods for IV selection in high-dimensional settings, recommending the de-biased approach when computational resources permit.

Building upon MR-SPLIT, we proposed MR-SPLIT+, which further relaxes the plurality rule to accommodate invalid IVs. By incorporating best subset selection and repeated sample splitting, MR-SPLIT+ achieves remarkable improvements in the accuracy of invalid IV identification, and demonstrates strong selection consistency in both theory and practice. Simulation results indicate that MR-SPLIT+ yields performance close to that of the oracle TSLS method and maintains high computational efficiency even in large samples. Although MR-SPLIT+ is robust, there remain opportunities for further generalization, such as handling binary exposures/outcomes and bidirectional causal relationships. The method also holds promise for constructing causal networks involving multiple exposures.

Finally, we extended our framework to bidirectional causal inference and developed BiMR-

SPLIT+, which enables robust identification of causal effects in the presence of invalid IVs and bidirectional causality. BiMR-SPLIT+ offers enhanced computational efficiency over MR-SPLIT+ and demonstrates the lowest estimation bias among competing methods (except for the oracle) while maintaining high coverage probabilities. Its efficacy was validated both in simulations and in a Drosophila gene expression application, with findings consistent with known biological mechanisms.

Collectively, these methodological advances significantly improve the reliability and applicability of MR for complex causal inference tasks. Our frameworks provide practical solutions to weak and invalid IV problems, enable robust bidirectional inference, and offer new avenues for future methodological developments, such as causal network construction and non-linear MR analysis.

Overall, the methods developed in this dissertation provide a unified and flexible framework for advancing robust causal inference in genetic epidemiology.

5.2 Biological Insights

Beyond methodological advances, the approaches developed in this dissertation have yielded valuable insights into important biological questions.

Applying MR-SPLIT to the CRIC dataset, we established a robust causal relationship between kidney function—as measured by eGFR and uACR—and apparent treatment-resistant hypertension (aTRH). These findings are consistent with prior observational and clinical evidence, but our methods provide stronger statistical support by effectively controlling for weak and invalid instruments. The use of debiased IV selection further enhanced the reliability of these conclusions in the high-dimensional genomic context.

In large-scale analyses with UK Biobank data, MR-SPLIT+ was validated using both positive and negative control designs. The method not only reproduced known causal effects, such as that of BMI on diastolic blood pressure, but also demonstrated robustness by correctly identifying the absence of implausible causal relationships, such as from adult BMI to birth weight. These results underscore the improved reliability and specificity of MR-SPLIT+ for causal inference in complex, real-world genomic studies.

Furthermore, the extension to bidirectional MR, as implemented in BiMR-SPLIT+, enabled us to disentangle complex causal relationships between gene expression and behavioral traits in *Drosophila melanogaster*. Specifically, our application to phototaxis data identified gene expressions with significant causal effects on phototactic behavior in both young and aged fly cohorts. Many of the identified candidate genes were corroborated by existing biological literature, while others represent novel findings that warrant further investigation. These results underscore the potential of our methods to uncover previously unrecognized mechanisms underlying complex phenotypes.

Overall, the application of MR-SPLIT, MR-SPLIT+, and BiMR-SPLIT+ to real-world datasets demonstrates their ability to generate robust and biologically meaningful inferences. These insights not only validate known biological relationships but also highlight new avenues for functional genomics and translational research.

5.3 Limitations of the Study

Despite the contributions of this work, several limitations should be acknowledged.

The dataset used in this study is mostly derived from the UK Biobank, a large-scale prospective cohort study comprising over 500,000 participants aged over 40 years at recruitment. While the breadth and depth of phenotypic and genotypic data in UK Biobank provide valuable opportunities for epidemiological research, it is important to acknowledge the limitations arising from its non-random sampling strategy. Specifically, UK Biobank participants tend to be healthier, older, more educated, and of higher socioeconomic status compared to the general UK population. In this context, although our real data analysis yields internally valid estimates, the potential selection bias may limit the generalizability of our findings to broader or more diverse populations.

In Chapter 3, the mediation analysis was also limited by the restricted availability of biological variables in the dataset, which constrained the range of potential mediators. As a result, important biological pathways such as hormonal regulation, autonomic nervous system activity, or vascular remodeling could not be assessed. Future studies with richer and more comprehensive biomarker panels may enable a more complete understanding of the mechanisms linking adiposity to blood pressure.

In the real data application of Chapter 4, the limitation lies in the relatively small sample size. The largest usable sample we could identify consisted of approximately 200 individuals, which may limit statistical power and generalizability. Nonetheless, the analysis revealed several interesting and biologically plausible findings. We believe that as larger datasets become available, the proposed method has the potential to uncover more nuanced and complex causal relationships. On the other hand, our analysis focused exclusively on exploring causal effects from gene expressions to phenotypic traits, without accounting for potential causal relationships among gene expressions themselves. In complex biological systems, it is plausible that certain gene expression levels influence traits indirectly through regulatory interactions with other genes. Ignoring such upstream or intermediate transcriptional pathways may obscure a more complete understanding of the causal architecture. Future work could extend the current framework to model gene expression networks and disentangle direct and indirect effects along regulatory cascades.

5.4 Future Research Directions

As large-scale biobank resources and individual-level genomic data continue to grow, there is tremendous potential for applying BiMR-SPLIT+ and related methods to a wider range of biological questions. In particular, these methods can play an increasingly important role in uncovering bidirectional causal relationships between complex traits and diseases, and in mapping gene expression networks that drive disease risk. Such analyses could ultimately contribute to more precise identification of causal genetic factors, providing new leads for functional studies and potential drug targets.

On the methodological side, further work is needed to strengthen the theoretical underpinnings of these approaches. For example, establishing sharper theoretical guarantees, improving the accuracy of confidence interval construction, and better understanding the behavior of these estimators in various practical settings remain open problems. There is also considerable scope for extending the current framework to accommodate nonlinear relationships, as well as binary exposures and outcomes, so that the methods can be used in an even broader array of applications.

Beyond these directions, new challenges and opportunities will no doubt arise as richer datasets become available, including the integration of multiple omics layers and the development of robust sensitivity analyses to assess model assumptions. Continued collaboration between methodological and applied researchers will be essential to fully realize the potential of these approaches in advancing both fundamental biology and clinical research.

Taken together, I hope that the work presented in this dissertation will serve as a foundation for further methodological development and inspire future applications that deepen our understanding of complex disease mechanisms.

BIBLIOGRAPHY

Abdelilah-Seyfried, S., Chan, Y.-M., Zeng, C., Justice, N. J., Younger-Shepherd, S., Sharp, L. E., Barbel, S., Meadows, S. A., Jan, L. Y., and Jan, Y. N. (2000). A gain-of-function screen for genes that affect the development of the drosophila adult external sensory organ. *Genetics*, 155(2):733–752.

Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212.

Anderson, T. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, 127(1):1–16.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Angrist, J. D. and Krueger, A. B. (1999). Chapter 23 - empirical strategies in labor economics. volume 3 of *Handbook of Labor Economics*, pages 1277–1366. Elsevier.

Apfel, N. and Liang, X. (2024). Agglomerative hierarchical clustering for selecting valid instrumental variables. *Journal of Applied Econometrics*, 39(7):1201–1219.

Avidor-Reiss, T., Maer, A. M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., and Zuker, C. S. (2004). Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell*, 117(4):527–539.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340.

Barro, R. J. (1997). *Macroeconomics*. MIT Press.

Berridge, M. J. (1998). Neuronal calcium signaling. *Neuron*, 21(1):13–26.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813 – 852.

Blomquist, S. and Dahlberg, M. (1999). Small sample properties of liml and jackknife iv estimators: Experiments with weak instruments. *Journal of Applied Econometrics*, 14(1):69–88.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2):512–525.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314.

Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., and Davey Smith, G. (2019). Improving the accuracy of two-sample summary-data mendelian randomization: moving beyond the nome assumption. *International Journal of Epidemiology*, 48(3):728–742.

Brini, M., Calì, T., Ottolini, D., and Carafoli, E. (2014). Neuronal calcium signaling: function and dysfunction. *Cellular and Molecular Life Sciences*, 71:2787–2814.

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., et al. (2014). Diversity and dynamics of the drosophila transcriptome. *Nature*, 512(7515):393–399.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665.

Burgess, S., Davies, N. M., Thompson, S. G., Consortium, E.-I., et al. (2014). Instrumental variable analysis with a nonlinear exposure–outcome relationship. *Epidemiology*, 25(6):877–885.

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):376.

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26(5):2333–2355.

Burgess, S., Smith, G. D., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., Hartwig, F. P., Kutalik, Z., Holmes, M. V., Minelli, C., et al. (2019). Guidelines for performing mendelian randomization investigations: update for summer 2023. *Wellcome Open Research*, 4.

Burgess, S., Thompson, S. G., and Collaboration, C. C. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764.

Carboni, A. L., Hanson, M. A., Lindsay, S. A., Wasserman, S. A., and Lemaitre, B. (2022). Cecropins contribute to drosophila host defense against a subset of fungal and gram-negative bacterial infection. *Genetics*, 220(1):iyab188.

Castle, W. E. (1903). Mendel's law of heredity. *Science*, 18(456):396–406.

Chen, J., Bundy, J. D., Hamm, L. L., Hsu, C.-y., Lash, J., Miller III, E. R., Thomas, G., Cohen, D. L., Weir, M. R., Raj, D. S., et al. (2019). Inflammation and apparent treatment-resistant hypertension in patients with chronic kidney disease: the results from the cric study. *Hypertension*, 73(4):785–793.

Chen, S. (2025). Two-sample bi-directional causality between two traits with some invalid ivs in both directions using gwas summary statistics. *Human Genetics and Genomics Advances*, page 100449.

Cheng, B., Bai, Y., Liu, L., Meng, P., Cheng, S., Yang, X., Pan, C., Wei, W., Liu, H., Jia, Y., et al. (2024). Mendelian randomization study of the relationship between blood and urine biomarkers and schizophrenia in the uk biobank cohort. *Communications Medicine*, 4(1):40.

Ciurciu, A., Duncalf, L., Jonchere, V., Lansdale, N., Vasieva, O., Glenday, P., Rudenko, A., Vissi, E., Cobbe, N., Alphey, L., et al. (2013). Pnuts/pp1 regulates rnapii-mediated gene expression and is necessary for developmental growth. *PLOS Genetics*, 9(10):e1003885.

Coresh, J., Wei, G. L., McQuillan, G., Brancati, F. L., Levey, A. S., Jones, C., and Klag, M. J. (2001). Prevalence of high blood pressure and elevated serum creatinine level in the united states: findings from the third national health and nutrition examination survey (1988-1994). *Archives of Internal Medicine*, 161(9):1207–1216.

Das, R., Sebo, Z., Pence, L., and Dobens, L. L. (2014). Drosophila tribbles antagonizes insulin signaling-mediated growth and metabolism via interactions with akt kinase. *PLOS One*, 9(10):e109530.

Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*, 32(1):1–22.

Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98.

Davies, S.-A., Overend, G., Sebastian, S., Cundall, M., Cabrero, P., Dow, J. A., and Terhzaz, S. (2012). Immune and stress response 'cross-talk'in the drosophila malpighian tubule. *Journal of Insect Physiology*, 58(4):488–497.

Denault, W. R. P., Bohlin, J., Page, C. M., Burgess, S., and Jugessur, A. (2022). Cross-fitted instrument: A blueprint for one-sample mendelian randomization. *PLOS Computational Biology*, 18(8):1–21.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical Science*, pages 533–558.

Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69(5):1161–1191.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Ferro, E. and Trabalzini, L. (2010). Ralgds family members couple ras to ral signalling and that's not all. *Cellular Signalling*, 22(12):1804–1810.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 16(2):175–185.

Findlay, G. D., MacCoss, M. J., and Swanson, W. J. (2009). Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in drosophila. *Genome Research*, 19(5):886–896.

Gene Ontology Curators (2002–). Manual transfer of experimentally-verified manual GO annotation data to orthologs by curator judgment of sequence similarity. Personal communication to FlyBase, FlyBase ID: FBrf0255270.

Gerber, M., Eissenberg, J. C., Kong, S., Tenney, K., Conaway, J. W., Conaway, R. C., and Shilatifard, A. (2004). In vivo requirement of the rna polymerase ii elongation factor elongin a for proper gene expression and development. *Molecular and Cellular Biology*, 24(22):9911–9919.

Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012). Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339.

Gordon, M. D., Ayres, J. S., Schneider, D. S., and Nusse, R. (2008). Pathogenesis of listeria-infected drosophila wntd mutants is associated with elevated levels of the novel immunity gene edin. *PLOS Pathogens*, 4(7):e1000111.

Greco M, F. D., Minelli, C., Sheehan, N. A., and Thompson, J. R. (2015). Detecting pleiotropy in mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine*, 34(21):2926–2940.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):793–815.

Guo, Z. and Small, D. S. (2016). Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 17(100):1–35.

Hartwig, F. P., Davies, N. M., Hemani, G., and Davey Smith, G. (2016). Two-sample mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique.

Hartwig, F. P., Tilling, K., Davey Smith, G., Lawlor, D. A., and Borges, M. C. (2021). Bias in two-sample mendelian randomization when using heritable covariable-adjusted summary associations. *International Journal of Epidemiology*, 50(5):1639–1650.

He, Q., Ding, Z. Y., Fong, D. Y.-T., and Karlberg, J. (2000). Blood pressure is associated with body mass index in both normal and obese children. *Hypertension*, 36(2):165–170.

He, Z., Chen, Z., de Borst, M. H., Zhang, Q., of Blood Pressure, I. C., Snieder, H., and Thio, C. H. (2023). Observational and genetic evidence for bidirectional effects between red blood cell traits and diastolic blood pressure. *American Journal of Hypertension*, 36(10):551–560.

Hemani, G., Bowden, J., and Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human Molecular Genetics*, 27(R2):R195–R208.

Impey, S., Obrietan, K., and Storm, D. R. (1999). Making new connections: role of erk/map kinase signaling in neuronal plasticity. *Neuron*, 23(1):11–14.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561.

Jiang, T., Gill, D., Butterworth, A. S., and Burgess, S. (2023). An empirical investigation into the impact of winner's curse on estimates from mendelian randomization. *International Journal of Epidemiology*, 52(4):1209–1219.

Judd, E. and Calhoun, D. (2014). Apparent and true resistant hypertension: definition, prevalence and outcomes. *Journal of Human Hypertension*, 28(8):463–468.

Kaboré, J., Metzger, M., Helmer, C., Berr, C., Tzourio, C., Drueke, T. B., Massy, Z. A., and Stengel, B. (2017). Hypertension control, apparent treatment resistance, and outcomes in the elderly population with chronic kidney disease. *Kidney International Reports*, 2(2):180–191.

Kabore, J., Metzger, M., Helmer, C., Berr, C., Tzourio, C., Massy, Z. A., and Stengel, B. (2016). Kidney function decline and apparent treatment-resistant hypertension in the elderly. *PLOS One*, 11(1):e0146056.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.

Katoh, T. and Tiemeyer, M. (2013). The n's and o's of drosophila glycoprotein glycobiology. *Glycoconjugate Journal*, 30:57–66.

Khandaker, G. M., Pearson, R. M., Zammit, S., Lewis, G., and Jones, P. B. (2014). Association of serum interleukin 6 and c-reactive protein in childhood with depression and psychosis in young adult life: a population-based longitudinal study. *JAMA Psychiatry*, 71(10):1121–1128.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803.

Kolesár, M. (2018). Minimum distance approach to inference with many instruments. *Journal of Econometrics*, 204(1):86–100.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484.

Lattmann, S., Giri, B., Vaughn, J. P., Akman, S. A., and Nagamine, Y. (2010). Role of the amino terminal rhau-specific motif in the recognition and resolution of guanine quadruplex-rna by the deah-box rna helicase rhau. *Nucleic Acids Research*, 38(18):6219–6233.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.

Li, C. (2019). *Rethinking Nonlinear Instrumental Variables*. PhD thesis, Duke University.

Lin, Y., Windmeijer, F., Song, X., and Fan, Q. (2024). On the instrumental variable estimation with many weak and invalid instruments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):1068–1088.

Linderman, G. C., Lu, J., Lu, Y., Sun, X., Xu, W., Nasir, K., Schulz, W., Jiang, L., and Krumholz, H. M. (2018). Association of body mass index with blood pressure among 1.7 million chinese adults. *JAMA Network Open*, 1(4):e181271–e181271.

Liu, Y. and Xie, J. (2019). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115:1–29.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

Maina, J. G., Balkhiyarova, Z., Nouwen, A., Pupko, I., Ulrich, A., Boissel, M., Bonnefond, A., Froguel, P., Khamis, A., Prokopenko, I., et al. (2023). Bidirectional mendelian randomization and multiphenotype gwas show causality and shared pathophysiology between depression and type 2 diabetes. *Diabetes Care*, 46(9):1707–1714.

Mammen, G. and Faulkner, G. (2013). Physical activity and the prevention of depression: a systematic review of prospective studies. *American Journal of Preventive Medicine*, 45(5):649–657.

Marmot, M. and Brunner, E. (1991). Alcohol and cardiovascular disease: the status of the u shaped curve. *BMJ: British Medical Journal*, 303(6802):565.

Members, I. P. (2004–). Gene ontology annotation through association of interpro records with go terms. FlyBase ID: FBrf0174215.

Millard, L. A., Davies, N. M., Tilling, K., Gaunt, T. R., and Davey Smith, G. (2019). Searching for the causal effects of body mass index in over 300 000 participants in uk biobank, using mendelian randomization. *PLOS Genetics*, 15(2):e1007951.

Miller, A. (2002). *Subset selection in regression*. Chapman and Hall/CRC.

Minelli, C., Del Greco M, F., van der Plaat, D. A., Bowden, J., Sheehan, N. A., and Thompson, J. (2021). The use of two-sample methods for mendelian randomization analyses on single large datasets. *International Journal of Epidemiology*, 50(5):1651–1659.

Mukherjee, S., Paricio, N., and Sokol, N. S. (2021). A stress-responsive mirna regulates bmp signaling to maintain tissue homeostasis. *Proceedings of the National Academy of Sciences*, 118(21):e2022583118.

Müller, R., Hülsmeier, A. J., Altmann, F., Ten Hagen, K., Tiemeyer, M., and Hennet, T. (2005). Characterization of mucin-type core-1 β1-3 galactosyltransferase homologous enzymes in drosophila melanogaster. *The FEBS Journal*, 272(17):4295–4305.

Nászai, M., Bellec, K., Yu, Y., Roman-Fernandez, A., Sandilands, E., Johansson, J., Campbell, A. D., Norman, J. C., Sansom, O. J., Bryant, D. M., et al. (2021). Ral gtpases mediate egfr-driven intestinal stem cell proliferation and tumourigenesis. *Elife*, 10:e63807.

Panagiotou, O. A., Ioannidis, J. P. A., and for the Genome-Wide Significance Project (2011). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1):273–286.

Patel, A., DiTraglia, F. J., Zuber, V., and Burgess, S. (2024). Selecting invalid instruments to improve mendelian randomization with two-sample summary data. *The Annals of Applied Statistics*, 18(2):1729–1749.

Pierce, B. L., Ahsan, H., and VanderWeele, T. J. (2010). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology*, 40(3):740–752.

Project, G. R. G. (2011). Phylogenetic annotation using the gene ontology. Personal communication to FlyBase. FlyBase ID: FBrf0258542.

Qi, G. and Chatterjee, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications*, 10(1):1941.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415.

Schuch, F. B., Vancampfort, D., Firth, J., Rosenbaum, S., Ward, P. B., Silva, E. S., Hallgren, M., Ponce De Leon, A., Dunn, A. L., Deslandes, A. C., et al. (2018). Physical activity and incident depression: a meta-analysis of prospective cohort studies. *American Journal of Psychiatry*, 175(7):631–648.

Shea, J. (1997). Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics*, 79(2):348–352.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232.

Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832.

Shi, R., Wang, L., Burgess, S., and Cui, Y. (2024). Mr-split: a novel method to address selection and weak instrument bias in one-sample mendelian randomization studies. *PLOS Genetics*, 20(9):e1011391.

Shi, R., Wang, L., and Cui, Y. (2025). Mr-split+: a unified method for many weak and invalid instruments with selection bias control in one-sample mendelian randomization studies. Manuscript submitted for publication.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312.

Sproviero, W., Winchester, L., Newby, D., Fernandes, M., Shi, L., Goodday, S. M., Prats-Uribe, A., Alhambra, D. P., Buckley, N. J., and Nevado-Holgado, A. J. (2021). High blood pressure and risk of dementia: a two-sample mendelian randomization study in the uk biobank. *Biological psychiatry*, 89(8):817–824.

Staiger, D. and Stock, J. H. (1994). Instrumental variables regression with weak instruments. Working Paper 151, National Bureau of Economic Research.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779.

Sulc, J., Sjaarda, J., and Kutalik, Z. (2022). Polynomial mendelian randomization reveals non-linear causal effects for obesity-related traits. *Human Genetics and Genomics Advances*, 3(3).

Thomas, D. C., Lawlor, D. A., and Thompson, J. R. (2007). Re: Estimation of bias in nongenetic observational studies using" mendelian triangulation" by bautista et al.

Thomas, G., Xie, D., Chen, H.-Y., Anderson, A. H., Appel, L. J., Bodana, S., Brecklin, C. S., Drawz, P., Flack, J. M., Miller III, E. R., et al. (2016). Prevalence and prognostic significance of apparent treatment resistant hypertension in chronic kidney disease: report from the chronic renal insufficiency cohort study. *Hypertension*, 67(2):387–396.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Torroja, L., Ortuño-Sahagún, D., Ferrús, A., Hämmerle, B., and Barbas, J. A. (1998). scully, an essential gene of drosophila, is homologous to mammalian mitochondrial type ii l-3-hydroxyacyl-coa dehydrogenase/amyloid-β peptide-binding protein. *The Journal of Cell Biology*, 141(4):1009–1017.

TRYSELIUS, Y., Samakovlis, C., Kimbrell, D. A., and Hultmark, D. (1992). Cecc, a cecropin gene expressed during metamorphosis in drosophila pupae. *European Journal of Biochemistry,*, 204(1):395–399.

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nature Genetics*, 50(5):693–698.

Verleyen, P., Baggerman, G., D'Hertog, W., Vierstraete, E., Husson, S. J., and Schoofs, L. (2006). Identification of new immune induced molecules in the haemolymph of drosophila melanogaster by 2d-nanolc ms/ms. *Journal of Insect Physiology*, 52(4):379–388.

Wainwright, K. et al. (2005). *Fundamental methods of mathematical economics*. McGraw-Hill.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

Wang, S. and Kang, H. (2022). Weak-instrument robust tests in two-sample summary-data mendelian randomization. *Biometrics*, 78(4):1699–1713.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37(5A):2178.

Windmeijer, F., Liang, X., Hartwig, F. P., and Bowden, J. (2021). The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):752–776.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.

Xing, G., Gan, G., Chen, D., Sun, M., Yi, J., Lv, H., Han, J., and Xie, W. (2014). Drosophila neuroligin3 regulates neuromuscular junction development and synaptic differentiation. *Journal of Biological Chemistry*, 289(46):31867–31877.

Yamanaka, N., Romero, N. M., Martin, F. A., Rewitz, K. F., Sun, M., O'Connor, M. B., and Léopold, P. (2013). Neuroendocrine control of drosophila larval light preference. *Science*, 341(6150):1113–1116.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82.

Ye, T., Liu, Z., Sun, B., and Tchetgen Tchetgen, E. (2024). Genius-mawii: for robust mendelian randomization with many weak invalid instruments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):1045–1067.

You, H., Lattmann, S., Rhodes, D., and Yan, J. (2017). Rhau helicase stabilizes g4 in its nucleotide-free state and destabilizes g4 upon atp hydrolysis. *Nucleic Acids Research*, 45(1):206–214.

Yu, Z., Coresh, J., Qi, G., Grams, M., Boerwinkle, E., Snieder, H., Teumer, A., Pattaro, C., Köttgen, A., Chatterjee, N., et al. (2020). A bidirectional mendelian randomization study supports causal effects of kidney function on blood pressure. *Kidney international*, 98(3):708–716.

Yuan, Z., Liu, L., Guo, P., Yan, R., Xue, F., and Zhou, X. (2022). Likelihood-based mendelian randomization analysis with automated instrument selection and horizontal pleiotropic modeling. *Science Advances*, 8(9):eabl5744.

Yusni, Y., Rahman, S., and Naufal, I. (2024). Positive correlation between body weight and body mass index with blood pressure in young adults. *Narra J*, 4(1):e533.

Zhang, F. (2006). *The Schur complement and its applications*, volume 4. Springer Science & Business Media.

Zhao, G., Lu, Z., Sun, Y., Kang, Z., Feng, X., Liao, Y., Sun, J., Zhang, Y., Huang, Y., and Yue, W. (2023). Dissecting the causal association between social or physical inactivity and depression: a bidirectional two-sample mendelian randomization study. *Translational Psychiatry*, 13(1):194.

Zhao, H. and Wang, T. (2020). Pe homeostasis rebalanced through mitochondria-er lipid exchange prevents retinal degeneration in drosophila. *PLOS Genetics*, 16(10):e1009070.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

APPENDIX A

SUPPLEMENTARY MATERIALS

A.1 Codes

The R codes for MR-SPLIT can be freely accessed at: https://github.com/RuxinShi/MR-SPLIT

The R code to implement MR-SPLIT+ is available at an anonymous GitHub repository (for peer review): https://anonymous.4open.science/r/MR_SPLIT_plus-CFCB.

A.2 Chapter 2

A.2.1 Proof of Theorem 1

Proof of Theorem For a given sample $\{X, Y, G\}$, the two stage IV model is defined as,

$$\begin{aligned} X &= G\alpha + \varepsilon_1 \\ Y &= X\beta + \varepsilon_2 \end{aligned} \tag{S1}$$

where $(\varepsilon_1, \varepsilon_2)' \sim N(0, \sigma^2(\begin{matrix} 1 & \rho \\ \rho & 1 \end{matrix}))$ and the correlation ρ reflects the degree of confounding effect.

Suppose we split the data into two parts, $I_1 = \{X_1, Y_1, G_1\}$, and $I_2 = \{X_2, Y_2, G_2\}$. Each subset has equal sample size $N/2$, where N is the total sample size. We first use sample I_1 to identify major and weak IVs, then use sample I_2 for causal inference. Suppose we have identified $p_1^{(1)}$ major IVs and $p_2^{(1)}$ weak IVs with the estimated effect size denoted as $\hat{\alpha}_1 = (\hat{\alpha}'_{1,M}, \hat{\alpha}'_{1,W})' \in \mathbb{R}^{p_1^{(1)}+p_2^{(1)}}$ when regressing exposure X_1 with the SNPs in G_1.

In sample I_2, MR-SPLIT combines the selected weak IVs into a new composite IV and uses it as an IV along with the major IVs:

$$\hat{G}_2 = (G_{2,M}, G_{2,W}\hat{\alpha}_{1,W}) \in \mathbb{R}^{\frac{N}{2}\times(p_1^{(1)}+1)} \tag{S2}$$

Then, we can apply the stage one of 2SLS in sample I_2 using these IVs and get the estimates of the exposure in sample I_2:

$$\hat{X}_2 = \hat{G}_2(\hat{G}'_2\hat{G}_2)^{-1}\hat{G}'_2X_2 = H_{\hat{G}_2}X_2,$$

where $H_X = X(X'X)^{-1}X'$ for any matrix X.

Similarly, we can also get the estimates of the exposure in sample I_1 by using sample I_2 to select the major and weak IVs:

$$\hat{X}_1 = \hat{G}_1(\hat{G}_1'\hat{G}_1)^{-1}\hat{G}_1'X_1 = H_{\hat{G}_1}X_1$$

Let $\hat{X} = \begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$. In stage two, we get the estimate of MR-SPLIT as

$$\begin{aligned}\hat{\beta} &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\ &= \beta + (X_1'H_{\hat{G}_1}X_1 + X_2'H_{\hat{G}_2}X_2)^{-1}(X_1'H_{\hat{G}_1}\varepsilon_{2,1} + X_2'H_{\hat{G}_2}\varepsilon_{2,2})\end{aligned}$$

and write $\varepsilon_2 = \begin{pmatrix} \varepsilon_{2,1} \\ \varepsilon_{2,2} \end{pmatrix}$.

For CFMR, it combines all selected IVs into a single IV. We use the subscript C to denote variables used in CFMR:

$$\hat{G}_{2,C} = (G_{2,M}\hat{\alpha}_{1,M}, G_{2,W}\hat{\alpha}_{1,W}) = G_2\hat{\alpha}_1 \in \mathbb{R}^{n\times 1} \tag{S3}$$

Similarly, in sample I_1, we combine G_1 and get:

$$\hat{G}_{1,C} = G_1\hat{\alpha}_2 \in \mathbb{R}^{n\times 1} \tag{S4}$$

For CFMR, let $\hat{G}_C = \begin{pmatrix} \hat{G}_{1,C} \\ \hat{G}_{2,C} \end{pmatrix}, X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Apply 2SLS on $\{X, Y, \hat{G}_C\}$ we get

$$\begin{aligned}\hat{\beta}_C &= (X'H_{\hat{G}_C}X)^{-1}X'H_{\hat{G}_C}Y \\ &= \beta + (X'H_{\hat{G}_C}X)^{-1}X'H_{\hat{G}_C}\varepsilon_2\end{aligned}$$

In the following, we will show that

$$var(\hat{\beta}) \leq var(\hat{\beta}_C)$$

where $\hat{\beta}$ denotes the estimate by MR-SPLIT. Since $var(\hat{\beta}) = (X_1' H_{\hat{G}_1} X_1 + X_2' H_{\hat{G}_2} X_2)^{-1}\sigma^2$, $var(\hat{\beta}_C) = (X' H_{\hat{G}_C} X)^{-1}\sigma^2$, to prove $var(\hat{\beta}) \leq var(\hat{\beta}_C)$, we need to show

$$\begin{aligned} & X_1' H_{\hat{G}_1} X_1 + X_2' H_{\hat{G}_2} X_2 \geq X' H_{\hat{G}_C} X \\ \iff \quad & X' \begin{pmatrix} H_{\hat{G}_1} & \\ & H_{\hat{G}_2} \end{pmatrix} X \geq X' H_{\hat{G}_C} X \\ \iff \quad & X' \left(\begin{pmatrix} H_{\hat{G}_1} & \\ & H_{\hat{G}_2} \end{pmatrix} - H_{\hat{G}_C} \right) X \geq 0 \end{aligned}$$

Hence, it is sufficient to show

$$\begin{pmatrix} H_{\hat{G}_1} & \\ & H_{\hat{G}_2} \end{pmatrix} - H_{\hat{G}_C} \succeq 0 \tag{S5}$$

where for any matrix X, $X \succeq 0$ means it is positive semi-definite.

Recall that $\hat{G}_C = \begin{pmatrix} \hat{G}_{1,C} \\ \hat{G}_{2,C} \end{pmatrix}$,

$$H_{\hat{G}_C} = \hat{G}_C (\hat{G}_C' \hat{G}_C)^{-1} \hat{G}_C' = \frac{1}{\hat{G}_{1,C}' \hat{G}_{1,C} + \hat{G}_{2,C}' \hat{G}_{2,C}} \begin{pmatrix} \hat{G}_{1,C}\hat{G}_{1,C}' & \hat{G}_{1,C}\hat{G}_{2,C}' \\ \hat{G}_{2,C}\hat{G}_{1,C}' & \hat{G}_{2,C}\hat{G}_{2,C}' \end{pmatrix}$$

Let $a = \hat{G}_{1,C}' \hat{G}_{1,C} + \hat{G}_{2,C}' \hat{G}_{2,C} \in \mathbb{R}$, it remains to show

$$\begin{pmatrix} H_{\hat{G}_1} - \frac{\hat{G}_{1,C}\hat{G}_{1,C}'}{a} & -\frac{\hat{G}_{1,C}\hat{G}_{2,C}'}{a} \\ -\frac{\hat{G}_{2,C}\hat{G}_{1,C}'}{a} & H_{\hat{G}_2} - \frac{\hat{G}_{2,C}\hat{G}_{2,C}'}{a} \end{pmatrix} \succeq 0 \tag{S6}$$

From Eq. S2 - S4, we can get

$$\begin{aligned}
H_{\hat{G}_2} =&\begin{pmatrix} G_{2,M} & G_{2,W}\hat{\alpha}_{1,W} \end{pmatrix} \left(\hat{G}_2'\hat{G}_2\right)^{-1} \begin{pmatrix} G_{2,M}' \\ \hat{\alpha}_{1,W}'G_{2,W}' \end{pmatrix} \\
=&(\begin{matrix} G_{2,M} & G_{2,W}\hat{\alpha}_{1,W} \end{matrix}) \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} \begin{pmatrix} G_{2,M}' \\ \hat{\alpha}_{1,W}'G_{2,W}' \end{pmatrix} \\
=&G_{2,M}A_2G_{2,M}' + G_{2,W}\hat{\alpha}_{1,W}C_2G_{2,M}' + G_{2,M}B_2\hat{\alpha}_{1,W}'G_{2,W}' + G_{2,W}\hat{\alpha}_{1,W}D_2\hat{\alpha}_{1,W}'G_{2,W}' \\
\frac{\hat{G}_{2,C}\hat{G}_{2,C}'}{a} =&\frac{1}{a}\begin{pmatrix} G_{2,M} & G_{2,W} \end{pmatrix} \begin{pmatrix} \hat{\alpha}_{1,M}\hat{\alpha}_{1,M}' & \hat{\alpha}_{1,M}\hat{\alpha}_{1,W}' \\ \hat{\alpha}_{1,W}\hat{\alpha}_{1,M}' & \hat{\alpha}_{1,W}\hat{\alpha}_{1,W}' \end{pmatrix} \begin{pmatrix} G_{2,M}' \\ G_{2,W}' \end{pmatrix} \\
=&\frac{1}{a}(G_{2,M}\hat{\alpha}_{1,M}\hat{\alpha}_{1,M}'G_{2,M}' + G_{2,W}\hat{\alpha}_{1,W}\hat{\alpha}_{1,M}'G_{2,M}' \\
&+ G_{2,M}\hat{\alpha}_{1,M}\hat{\alpha}_{1,W}'G_{2,W}' + G_{2,W}\hat{\alpha}_{1,W}\hat{\alpha}_{1,W}'G_{2,W}')
\end{aligned} \tag{S7}$$

Therefore,

$$\begin{aligned}
H_{\hat{G}_2} - \frac{\hat{G}_{2,C}\hat{G}_{2,C}'}{a} =&G_{2,M}\left(A_2 - \frac{\hat{\alpha}_{1,M}\hat{\alpha}_{1,M}'}{a}\right)G_{2,M}' + G_{2,W}\hat{\alpha}_{1,W}\left(C_2 - \frac{\hat{\alpha}_{1,M}'}{a}\right)G_{2,M}' \\
&+ G_{2,M}\left(B_2 - \frac{\hat{\alpha}_{1,M}}{a}\right)\hat{\alpha}_{1,W}'G_{2,W}' + G_{2,W}\hat{\alpha}_{1,W}(D_2 - \frac{1}{a})\hat{\alpha}_{1,W}'G_{2,W}' \\
=&\begin{pmatrix} G_{2,M} & G_{2,W}\hat{\alpha}_{1,W} \end{pmatrix} \begin{pmatrix} A_2 - \frac{\hat{\alpha}_{1,M}\hat{\alpha}_{1,M}'}{a} & B_2 - \frac{\hat{\alpha}_{1,M}}{a} \\ C_2 - \frac{\hat{\alpha}_{1,M}'}{a} & D_2 - \frac{1}{a} \end{pmatrix} \begin{pmatrix} G_{2,M}' \\ \hat{\alpha}_{1,W}'G_{2,W}' \end{pmatrix} \\
=&\hat{G}_2Q_4\hat{G}_2'
\end{aligned} \tag{S8}$$

Similarly,

$$\begin{aligned}
H_{\hat{G}_1} - \frac{\hat{G}_{1,C}\hat{G}_{1,C}'}{a} =&\begin{pmatrix} G_{1,M} & G_{1,W}\hat{\alpha}_{2,W} \end{pmatrix} \begin{pmatrix} A_1 - \frac{\hat{\alpha}_{2,M}\hat{\alpha}_{2,M}'}{a} & B_1 - \frac{\hat{\alpha}_{2,M}}{a} \\ C_1 - \frac{\hat{\alpha}_{2,M}'}{a} & D_1 - \frac{1}{a} \end{pmatrix} \begin{pmatrix} G_{1,M}' \\ \hat{\alpha}_{2,W}'G_{1,W}' \end{pmatrix} \\
=&\hat{G}_1Q_1\hat{G}_1'
\end{aligned} \tag{S9}$$

Easily, we can also get

$$
\begin{aligned}
-\frac{\hat{G}_{1,C}\hat{G}'_{2,C}}{a} &= -\frac{1}{a}\begin{pmatrix} G_{1,M} & G_{1,W}\hat{\alpha}_{2,W} \end{pmatrix}\begin{pmatrix} \hat{\alpha}_{2,M}\hat{\alpha}_{1,M} & \hat{\alpha}_{2,M} \\ \hat{\alpha}'_{1,M} & 1 \end{pmatrix}\begin{pmatrix} G'_{2,M} \\ \hat{\alpha}'_{1,W}G'_{2,W} \end{pmatrix} \\
&= -\frac{1}{a}\hat{G}_1 Q_2 \hat{G}'_2 \qquad \text{(S10)} \\
-\frac{\hat{G}_{2,C}\hat{G}'_{1,C}}{a} &= -\frac{1}{a}\begin{pmatrix} G_{2,M} & G_{2,W}\hat{\alpha}_{1,W} \end{pmatrix}\begin{pmatrix} \hat{\alpha}_{1,M}\hat{\alpha}_{2,M} & \hat{\alpha}_{1,M} \\ \hat{\alpha}'_{2,M} & 1 \end{pmatrix}\begin{pmatrix} G'_{1,M} \\ \hat{\alpha}'_{2,W}G'_{1,W} \end{pmatrix} \\
&= -\frac{1}{a}\hat{G}_2 Q_3 \hat{G}'_1 \qquad \text{(S11)}
\end{aligned}
$$

Apply Eq.S8 - S11 to Eq. S6, we have

$$
\begin{pmatrix} \hat{G}_1 & \hat{G}_2 \end{pmatrix}\begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix}\begin{pmatrix} \hat{G}'_1 \\ \hat{G}'_2 \end{pmatrix} \succeq 0 \qquad \text{(S12)}
$$

We now only need to show that

$$
\begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix} \succeq 0. \qquad \text{(S13)}
$$

We first show that

$$
Q_4 = \begin{pmatrix} A_2 - \frac{\hat{\alpha}_{1,M}\hat{\alpha}'_{1,M}}{a} & B_2 - \frac{\hat{\alpha}_{1,M}}{a} \\ C_2 - \frac{\hat{\alpha}'_{1,M}}{a} & D_2 - \frac{1}{a} \end{pmatrix} \succ 0. \qquad \text{(S14)}
$$

Recall that

$$
\begin{aligned}
\begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} &= (\hat{G}'_2\hat{G}_2)^{-1} \\
&= \begin{pmatrix} G'_{2,M}G_{2,M} & G'_{2,M}G_{2,W}\hat{\alpha}_{1,W} \\ \hat{\alpha}'_{1,W}G'_{2,W}G_{2,M} & \hat{\alpha}'_{1,W}G'_{2,W}G_{2,W}\hat{\alpha}_{1,W} \end{pmatrix}^{-1} \qquad \text{(S15)}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
A_2 &= (G'_{2,M}G_{2,M})^{-1} + \frac{(G'_{2,M}G_{2,M})^{-1}(G'_{2,M}G_{2,W}\hat{\alpha}_{1,W}\hat{\alpha}'_{1,W}G'_{2,W}G_{2,M})(G'_{2,M}G_{2,M})^{-1}}{\hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}},\\
B_2 &= -\frac{(G'_{2,M}G_{2,M})^{-1}G'_{2,M}G_{2,W}\hat{\alpha}_{1,W}}{\hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}},\\
C_2 &= -\frac{\hat{\alpha}'_{1,W}G'_{2,W}G_{2,W}(G'_{2,M}G_{2,M})^{-1}}{\hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}},\\
D_2 &= \frac{1}{\hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
D_2 - \frac{1}{a} =& \frac{1}{\hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}} - \frac{1}{\hat{G}'_{1,C}\hat{G}_{1,C} + \hat{G}'_{2,C}\hat{G}_{2,C}}\\
=& \frac{1}{\hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}}\\
&- \frac{1}{\hat{G}'_{1,C}\hat{G}_{1,C} + (G_{2,M}\hat{\alpha}_{1,M} + G_{2,W}\hat{\alpha}_{1,W})'(G_{2,M}\hat{\alpha}_{1,M} + G_{2,W}\hat{\alpha}_{1,W})}
\end{aligned}
$$

Since

$$
\begin{aligned}
&(G_{2,M}\hat{\alpha}_{1,M} + G_{2,W}\hat{\alpha}_{1,W})'(G_{2,M}\hat{\alpha}_{1,M} + G_{2,W}\hat{\alpha}_{1,W}) - \hat{\alpha}'_{1,W}G'_{2,W}(I-H_{G_{2,M}})G_{2,W}\hat{\alpha}_{1,W}\\
=&(G_{2,M}\hat{\alpha}_{1,M} + H_{G_{2,M}}G_{2,W}\hat{\alpha}_{1,W})'(G_{2,M}\hat{\alpha}_{1,M} + H_{G_{2,M}}G_{2,W}\hat{\alpha}_{1,W}) > 0, \qquad \text{(S16)}
\end{aligned}
$$

we get $D_2 - \frac{1}{a} > 0$.

To prove S14, it is sufficient to show (Zhang, 2006)

$$
\begin{aligned}
&A_2 - \frac{\hat{\alpha}_{1,M}\hat{\alpha}'_{1,M}}{a} - (B_2 - \frac{\hat{\alpha}_{1,M}}{a})(D_2 - \frac{1}{Q})^{-1}(C_2 - \frac{\hat{\alpha}'_{1,M}}{a}) \succ 0\\
\Longleftrightarrow&(A_2 - \frac{\hat{\alpha}_{1,M}\hat{\alpha}'_{1,M}}{a})(D_2 - \frac{1}{a}) - (B_2 - \frac{\hat{\alpha}_{1,M}}{a})(C_2 - \frac{\hat{\alpha}'_{1,M}}{a}) \succ 0\\
\Longleftrightarrow&(a - \frac{1}{D_2})\tilde{A_2}^{-1} - \hat{\alpha}_{1,M}\hat{\alpha}'_{1,M} + \tilde{A_2}^{-1}\tilde{B}_2\tilde{C}_2\tilde{A_2}^{-1} - \hat{\alpha}_{1,M}\tilde{C}_2\tilde{A_2}^{-1} - \tilde{A_2}^{-1}\tilde{B}_2\hat{\alpha}'_{1,M} \succ 0 \qquad \text{(S17)}
\end{aligned}
$$

where $\begin{pmatrix} \tilde{A}_2 & \tilde{B}_2 \\ \tilde{C}_2 & \tilde{D}_2 \end{pmatrix} = \hat{G}'_2\hat{G}_2$. The left side of Eq. S17 can be obtained as $(a - \frac{1}{D_2})\tilde{A_2}^{-1} + (\tilde{A_2}^{-1}B_2 - \hat{\alpha}_{1,M})(\tilde{A_2}^{-1}B_2 - \hat{\alpha}_{1,M})'$, which is easy to verify to be a positive definite matrix.

To prove Eq. S13, now we only need to prove

$$Q_1 - Q_2 Q_4^{-1} Q3 \succeq 0. \tag{S18}$$

From Eq. S14, we have

$$\begin{aligned}
Q_4^{-1} &= \left((\hat{G}_2'\hat{G}_2)^{-1} - \frac{1}{a} \begin{pmatrix} \hat{\alpha}_{1,M} \\ 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_{1,M}' & 1 \end{pmatrix} \right)^{-1} \\
&= \left((\hat{G}_2'\hat{G}_2)^{-1} - \frac{1}{a} b_2 b_2' \right)^{-1} \\
&= \hat{G}_2'\hat{G}_2 - \frac{\hat{G}_2'\hat{G}_2 b_2 b_2' \hat{G}_2' \hat{G}_2}{b_2'\hat{G}_2'\hat{G}_2 b_2 - a},
\end{aligned} \tag{S19}$$

where $b_2 = \left(\hat{\alpha}_{1,M}', 1\right)'$, and the third equation utilizes the Woodbury matrix identity. Similarly, let $b_1 = \left(\hat{\alpha}_{2,M}', 1\right)'$,

$$Q_1 = (\hat{G}_1'\hat{G}_1)^{-1} - \frac{1}{a} b_1 b_1', \tag{S20}$$

$$Q_2 = -\frac{1}{a} b_1 b_2', \tag{S21}$$

$$Q_3 = -\frac{1}{a} b_2 b_1'. \tag{S22}$$

Substituting Eq. S19 -S22 into Eq. S18, we get

$$\begin{aligned}
&(\hat{G}_1'\hat{G}_1)^{-1} - \frac{1}{a} b_1 b_1' - \frac{1}{a^2} b_1 b_2' (\hat{G}_2'\hat{G}_2 - \frac{\hat{G}_2'\hat{G}_2 b_2 b_2' \hat{G}_2'\hat{G}_2}{b_2'\hat{G}_2'\hat{G}_2 b_2 - a}) b_2 b_1' \succeq 0 \\
\Longleftrightarrow &(\hat{G}_1'\hat{G}_1)^{-1} - b_1 (\frac{1}{a} + \frac{1}{a^2} b_2' (\hat{G}_2'\hat{G}_2 - \frac{\hat{G}_2'\hat{G}_2 b_2 b_2' \hat{G}_2'\hat{G}_2}{b_2'\hat{G}_2'\hat{G}_2 b_2 - a}) b_2) b_1' \succeq 0 \\
\Longleftrightarrow &(\hat{G}_1'\hat{G}_1)^{-1} - b_1 (\frac{1}{a} + \frac{\hat{G}_{2,C}'\hat{G}_{2,C}}{a^2} - \frac{(\hat{G}_{2,C}'\hat{G}_{2,C})^2}{a^2 \hat{G}_{2,C}'\hat{G}_{2,C} - a^3}) b_1' \succeq 0 \\
\Longleftrightarrow &(\hat{G}_1'\hat{G}_1)^{-1} - \frac{1}{\hat{G}_{1,C}'\hat{G}_{1,C}} b_1 b_1' \succeq 0
\end{aligned} \tag{S23}$$

Eq. S23 has a very similar structure as Eq. S14. It can be written as

$$\begin{pmatrix} A_1 - \frac{\hat{\alpha}_{2,M}\hat{\alpha}_{2,M}'}{\hat{G}_{1,C}'\hat{G}_{1,C}} & B_1 - \frac{\hat{\alpha}_{2,M}}{\hat{G}_{1,C}'\hat{G}_{1,C}} \\ C_1 - \frac{\hat{\alpha}_{2,M}'}{\hat{G}_{1,C}'\hat{G}_{1,C}} & D_1 - \frac{1}{\hat{G}_{1,C}'\hat{G}_{1,C}} \end{pmatrix} \succeq 0, \tag{S24}$$

which can be easily verified. This completes the proof of the theorem. □

A.2.2 Results of selecting major IVs under different partial F values

Table A.1 shows the results of distinguishing major and weak IVs with different partial F thresholds. We also showed the results with the criterion of $F > 100$. As demonstrated, this is exceedingly conservative and is generally best avoided. Furthermore, we recognize that a heritability ($h^2 = 0.5$) is considerably high for many exposure traits in practical scenarios, representing situations that are relatively uncommon in reality.

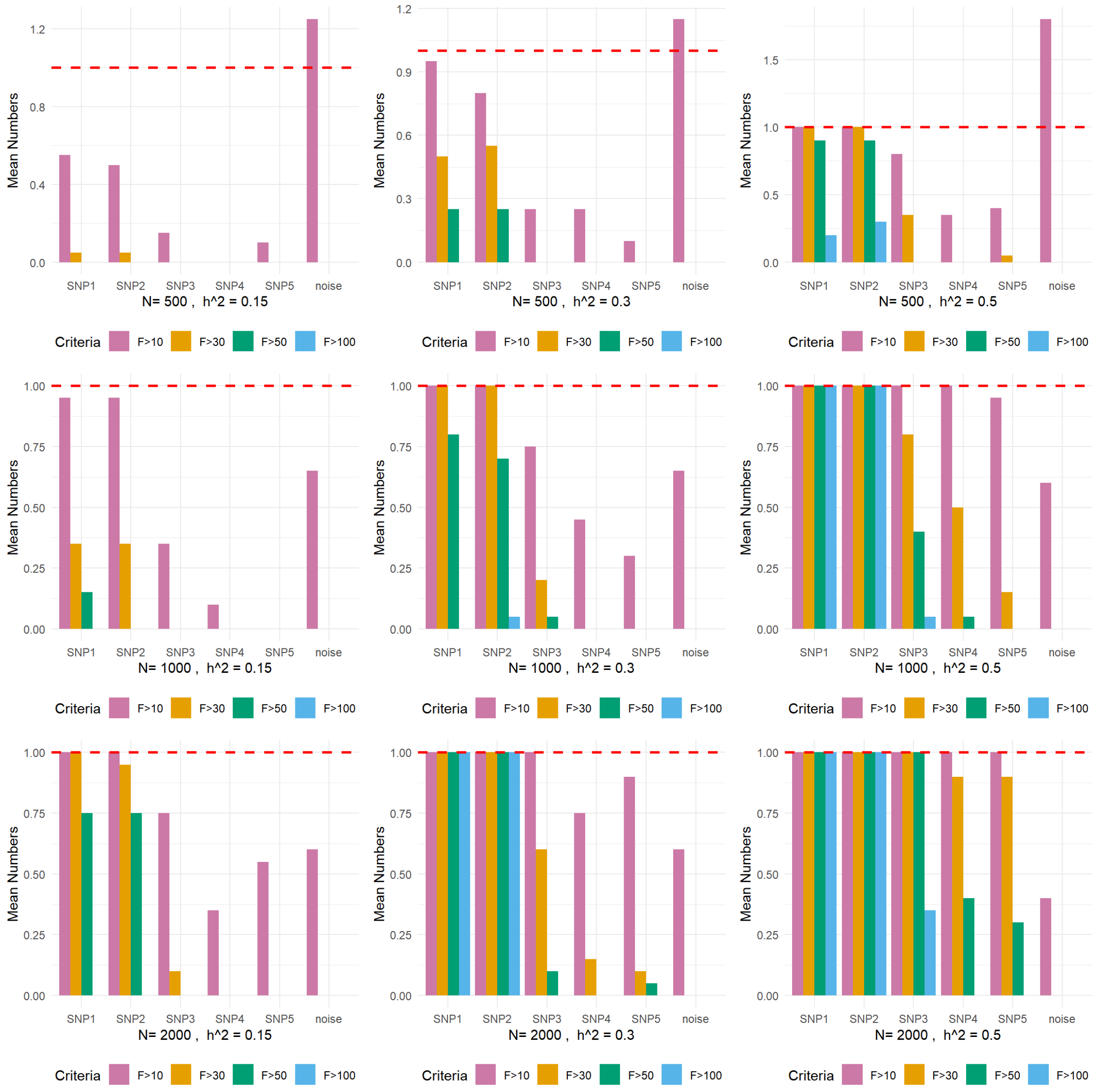


Figure A.1 Mean numbers of being identified as major IV using different thresholds in 1,000 simulations.

Table A.1 Mean numbers of being identified as major IV using different criteria in 1000 simulations. For the noise category, it was aggregated over the 295 null IVs.

h^2	N	Criteria	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	Noises (×295)
0.15	500	F>10	0.55	0.5	0.15	0	0.1	1.25
		F>30	0.05	0.05	0	0	0	0
		F>50	0	0	0	0	0	0
		F>100	0	0	0	0	0	0
	1000	F>10	0.95	0.95	0.35	0.1	0	0.65
		F>30	0.35	0.35	0	0	0	0
		F>50	0.15	0	0	0	0	0
		F>100	0	0	0	0	0	0
	2000	F>10	1	1	0.75	0.35	0.55	0.6
		F>30	1	0.95	0.1	0	0	0
		F>50	0.75	0.75	0	0	0	0
		F>100	0	0	0	0	0	0
0.3	500	F>10	0.95	0.8	0.25	0.25	0.1	1.15
		F>30	0.5	0.55	0	0	0	0
		F>50	0.25	0.25	0	0	0	0
		F>100	0	0	0	0	0	0
	1000	F>10	1	1	0.75	0.45	0.3	0.65
		F>30	1	1	0.2	0	0	0
		F>50	0.8	0.7	0.05	0	0	0
		F>100	0	0.05	0	0	0	0
	2000	F>10	1	1	1	0.75	0.9	0.6
		F>30	1	1	0.6	0.15	0.1	0
		F>50	1	1	0.1	0	0.05	0
		F>100	1	1	0	0	0	0
0.5	500	F>10	1	1	0.8	0.35	0.4	1.8
		F>30	1	1	0.35	0	0.05	0
		F>50	0.9	0.9	0	0	0	0
		F>100	0.2	0.3	0	0	0	0
	1000	F>10	1	1	1	1	0.95	0.6
		F>30	1	1	0.8	0.5	0.15	0
		F>50	1	1	0.4	0.05	0	0
		F>100	1	1	0.05	0	0	0
	2000	F>10	1	1	1	1	1	0.4
		F>30	1	1	1	0.9	0.9	0
		F>50	1	1	1	0.4	0.3	0
		F>100	1	1	0.35	0	0	0

A.2.3 Boxplots of causal effect estimates for MR-SPLIT, 2SLS, and LIML out of 1000 simulation runs under different scenarios.

LIML and 2SLS both use half of the dataset to select IVs and the other half to get the estimation. In contrast, LIML_w and 2SLS_w use the whole dataset for IV selection and causal effect estimation. When using half data to select IVs and another half for causal estimation, both MR-SPLIT and LIML produce unbiased estimates, though the variance for LIML is larger than MR-SPLIT. However, when using the whole data for both IV selection and causal effect estimation, LIML_w and 2SLS_w generate biased causal effect estimation. In either case, 2SLS yields biased effect estimates. This simulation demonstrates the issue of IV selection bias if it is not properly addressed.

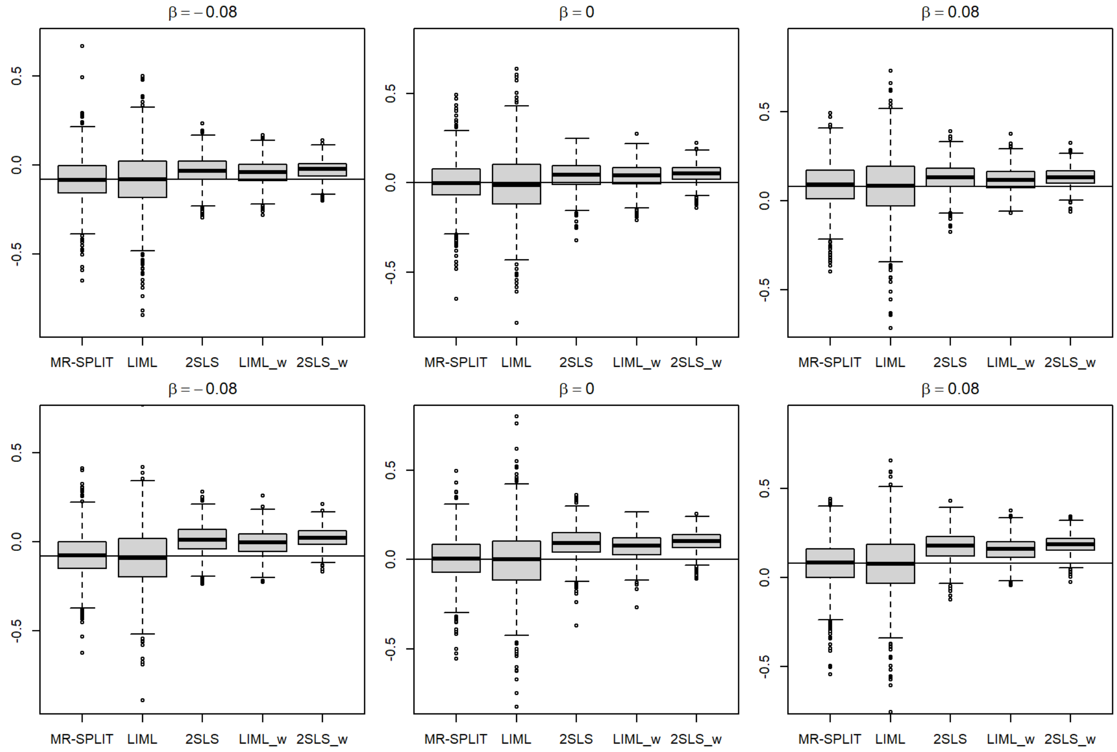


Figure A.2 Boxplots of causal effect estimates ($\hat{\beta}$) under $h^2 = 0.15$ and confounding correlation $\rho = 0.1$ (top) and $\rho = 0.2$ (bottom).

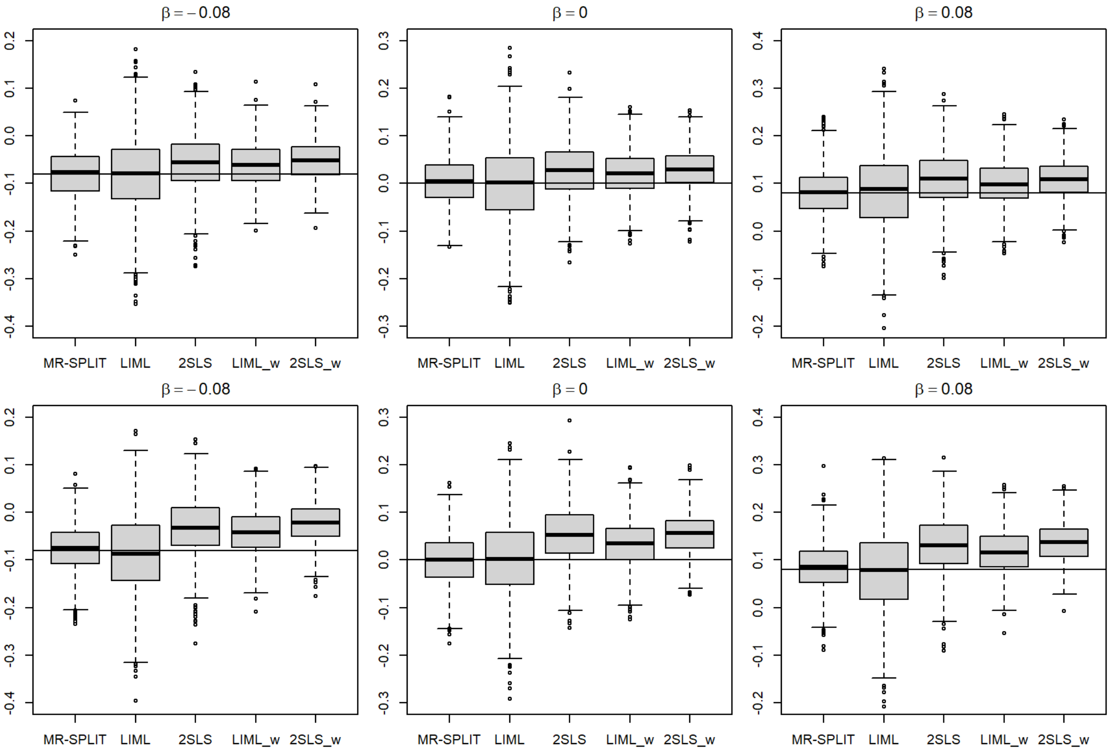


Figure A.3 Boxplots of causal effect estimates ($\hat{\beta}$) under $h^2 = 0.3$ and confounding correlation $\rho = 0.1$ (top) and $\rho = 0.2$ (bottom).

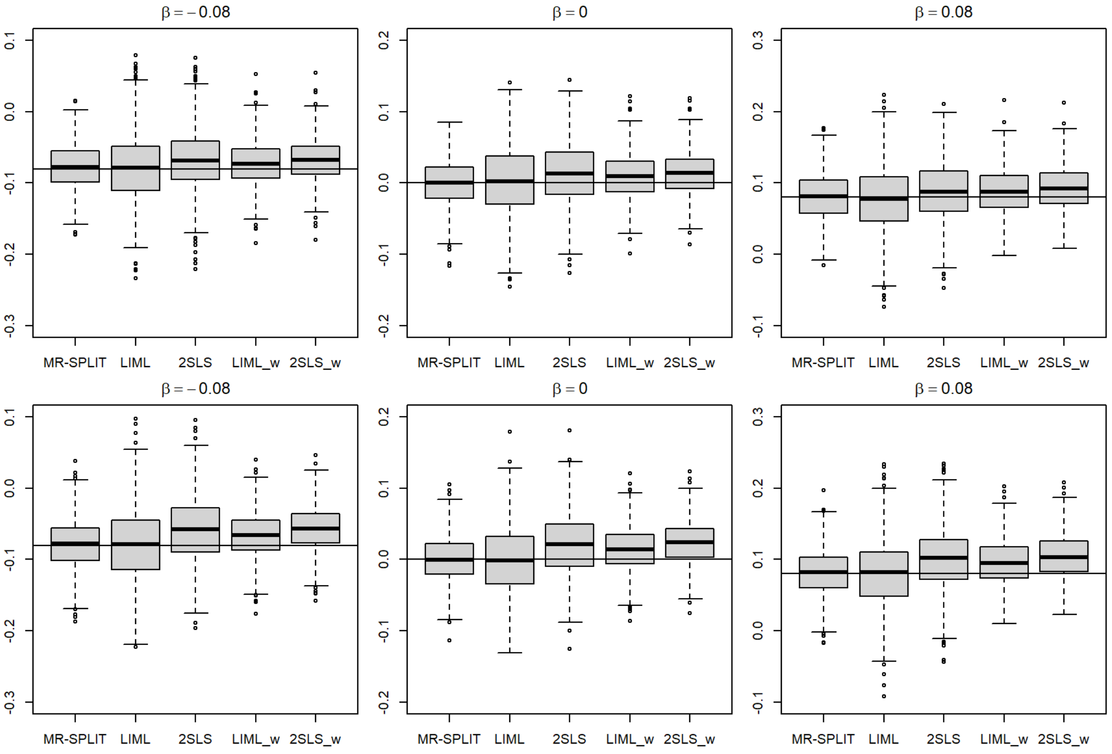


Figure A.4 Boxplots of causal effect estimates ($\hat{\beta}$) under $h^2 = 0.5$ and confounding correlation $\rho = 0.1$ (top) and $\rho = 0.2$ (bottom).

A.2.4 Boxplots of causal effect estimates for MR-SPLIT and CFMR out of 1000 simulation runs under different scenarios.

In nearly all scenarios, both CFMR and MR-SPLIT obtained approximately unbiased estimates. However, it is evident that MR-SPLIT consistently exhibits a smaller variance compared to CFMR. This can also be seen in the comparison of RMSE (see Figure A.11), where the RMSE of MR-SPLIT is always noticeably smaller than that of CFMR.

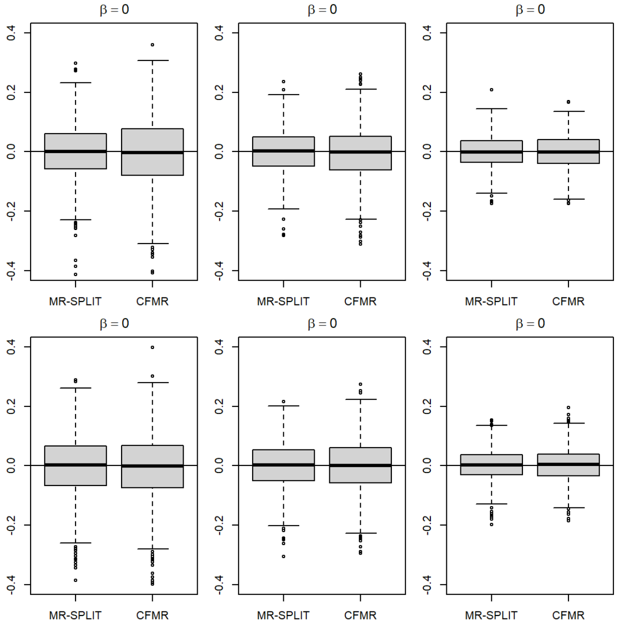


Figure A.5 Boxplots of causal effect estimates ($\hat{\beta}$) when $h^2 = 0.15$ (left), 0.2 (middle) , 0.3 (right) and sample size $N = 1000$ in scenario I (top) and scenario II (bottom).

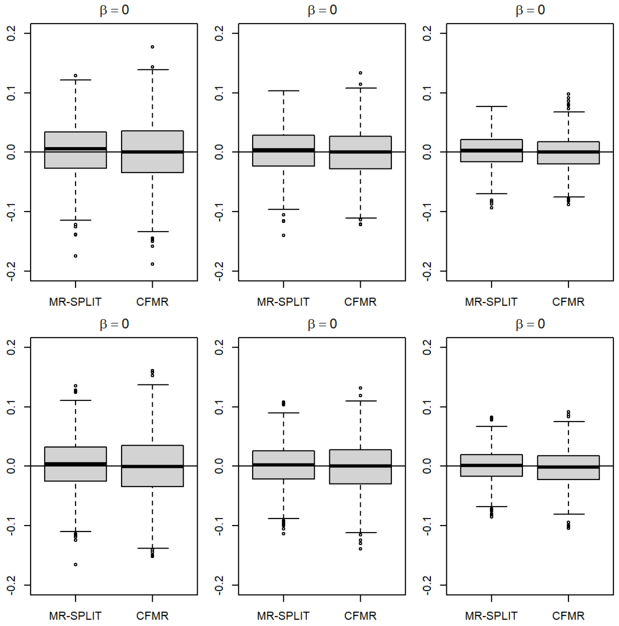


Figure A.6 Boxplots of causal effect estimates ($\hat{\beta}$) when h^2 = 0.15 (left), 0.2 (middle) , 0.3 (right) and sample size N = 3000 in scenario I (top) and scenario II (bottom).

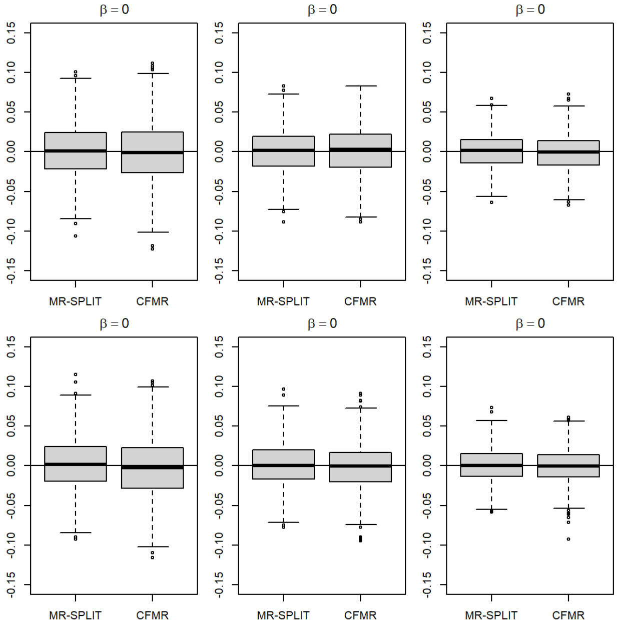


Figure A.7 Boxplots of causal effect estimates ($\hat{\beta}$) when h^2 = 0.15 (left), 0.2 (middle) , 0.3 (right) and sample size N = 5000 in scenario I (top) and scenario II (bottom).

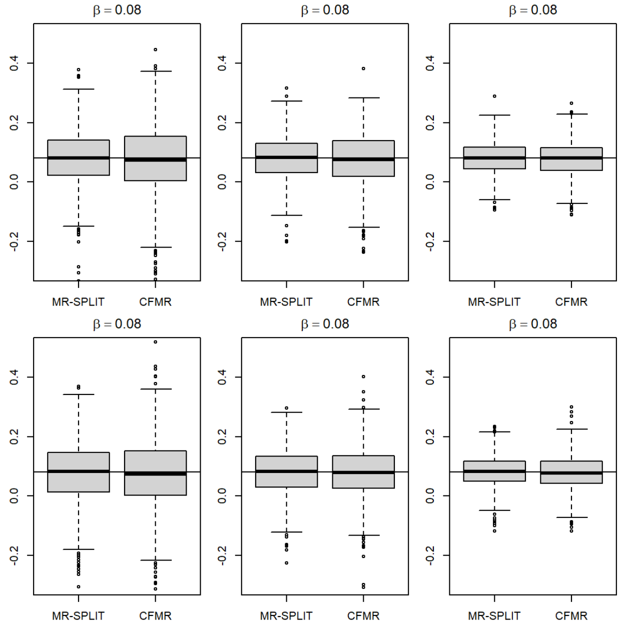


Figure A.8 Boxplots of causal effect estimates ($\hat{\beta}$) when h^2 = 0.15 (left), 0.2 (middle) , 0.3 (right) and sample size N = 1000 in scenario I (top) and scenario II (bottom).

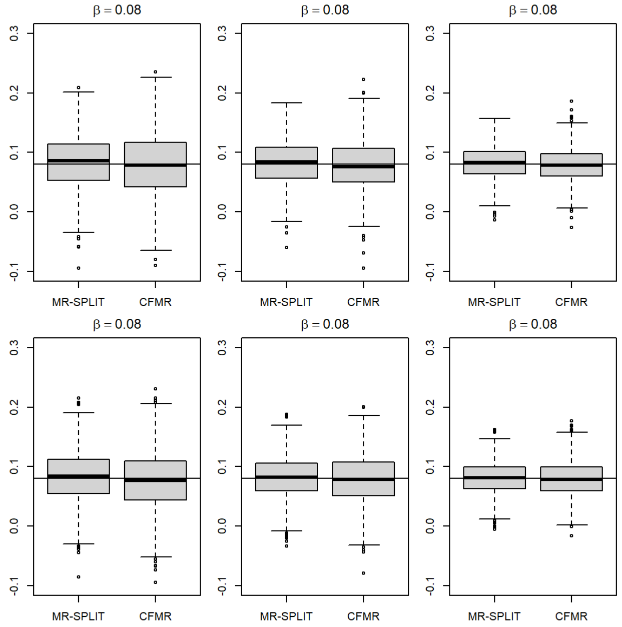


Figure A.9 Boxplots of causal effect estimates ($\hat{\beta}$) when h^2 = 0.15 (left), 0.2 (middle) , 0.3 (right) and sample size N = 3000 in scenario I (top) and scenario II (bottom).

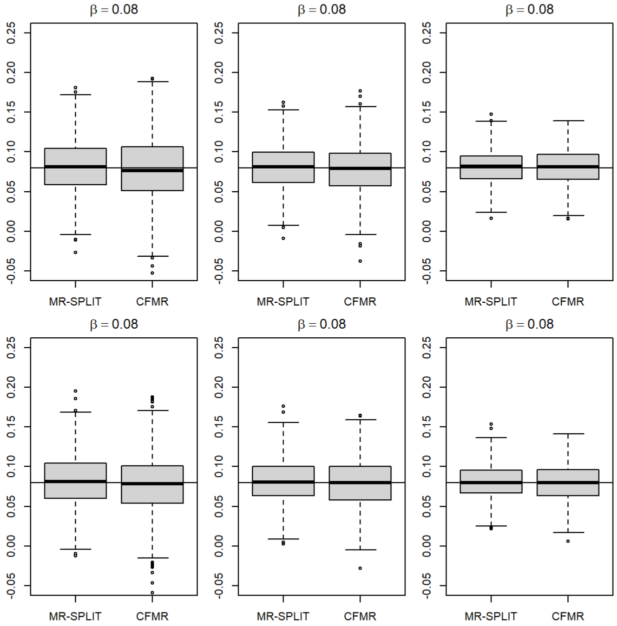


Figure A.10 Boxplots of causal effect estimates ($\hat{\beta}$) when h^2 = 0.15 (left), 0.2 (middle) , 0.3 (right) and sample size N = 5000 in scenario I (top) and scenario II (bottom).

A.2.5 RMSE comparison between MR-SPLIT and CFMR out of 1000 simulation runs under different scenarios.

The RMSE of MR-SPLIT is always smaller than that of CFMR, especially under a small sample size (e.g., $N = 1000$), indicating the estimation efficiency and consistency of MR-SPLIT compared to CFMR.

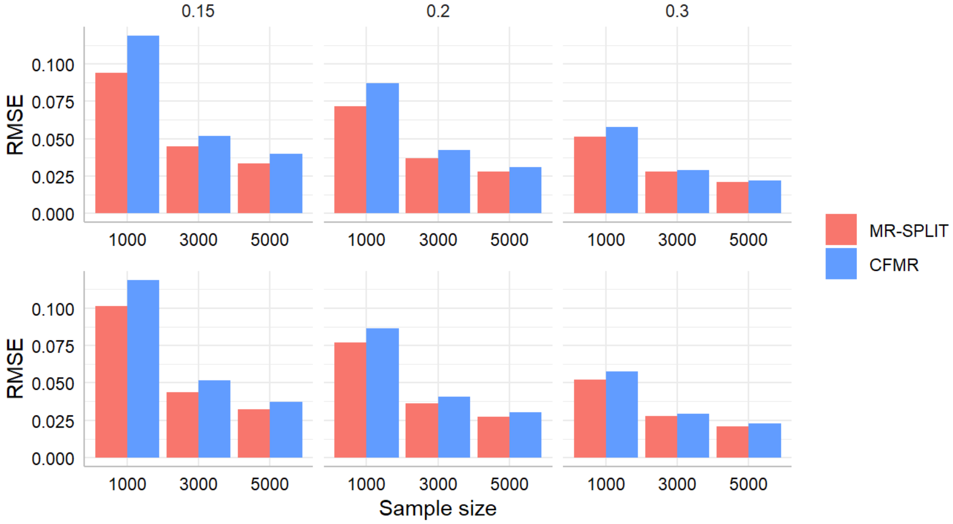


Figure A.11 RNSE comparison between MR-SPLIT and CFMR in Scenario I (top) and II (bottom).

A.2.6 Additional type I error and power simulation results for the evaluation of multiple data splitting

Figure A.12 shows the type I error out of 50 sample splits under $h^2 = 0.3$ and different sample sizes. When SNP heritability is significant and the sample size is relatively large (for instance, greater than 1000), the type I error stabilizes, even with a minimal number of sample splits.

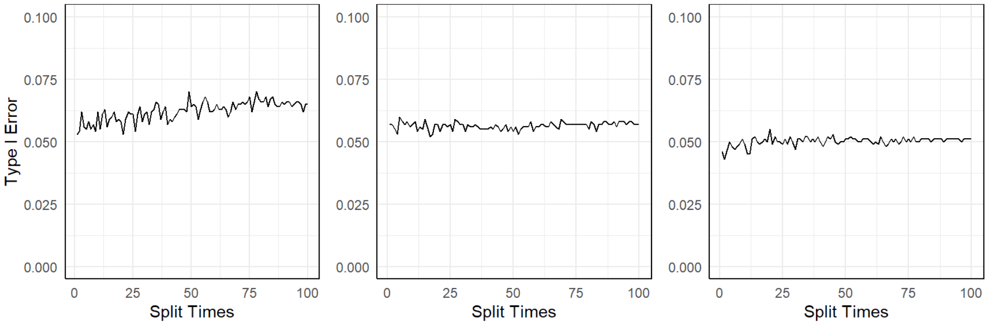


Figure A.12 Type I error when $h^2 = 0.3$ and $N = 500$ (left), 1000 (middle), 2000 (right) out of 50 sample splits.

Figure A.13 displays the empirical power under different sample sizes when $h^2 = 0.3$. Compared to scenarios where $h^2 = 0.15$ or 0.2, fewer splits are required to achieve optimal power. When the sample size is relatively small, for instance, $N = 500$, the power stabilizes after about 25 sample splits. As the sample size increases to 1000, a few sample splits are good enough to achieve stable power. The results suggest that in practice, one can lower the number of sample slits if the estimated SNP heritability for the exposure is strong and the sample size is large, to save computational time.

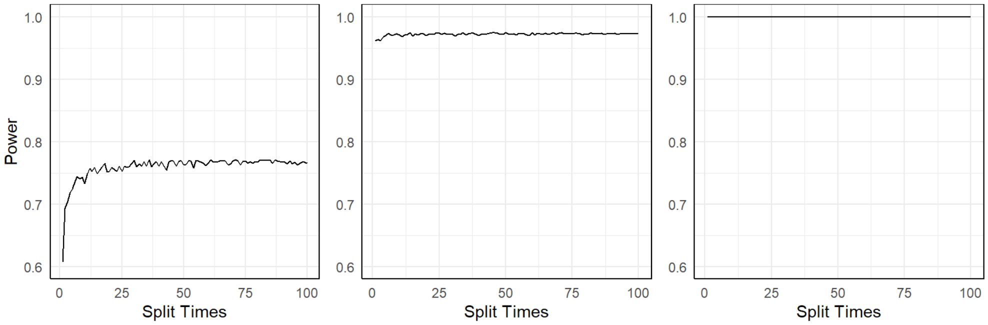


Figure A.13 Power performance when $h^2 = 0.3$ and $N = 500$ (left), 1000 (middle), 2000 (right) out of 50 sample splits.

A.2.7 Comparison between the LASSO and LASSO-projection methods.

In this simulation, we used two different methods, LASSO and LASSO-projection, to do the IV selection. We considered the case with the sample size as $N = 1,000$, and randomly generated a set of 300 independent SNPs with their minor allele frequency fixed as 0.3 for all the SNPs. We randomly chose 5 SNP IVs to generate the exposure variable. The correlation between the error terms is set to 0.16. The variation in the exposure explained by the 5 SNPs is set to $h^2 = 0.2$. The real causal effects β were set to $\{-0.08, 0, 0.08\}$. Both methods produced very similar effect estimates as revealed by the boxplots in Figure A.14. The LASSO-projection method yielded smaller type I error (Figure A.15, slightly larger power (Figure A.16 and smaller RMSE (Figure A.17), compared to the regular LASSO method. We also observed that the total number of IVs selected by the regular LASSO method is much larger than the LASSO-projection method (Figure A.18), and the number of major IVs selected by the LASSO-projection method is slightly higher than the LASSO method (Figure A.19). We observed similar trends under other settings and hence only reported the results of this scenario.

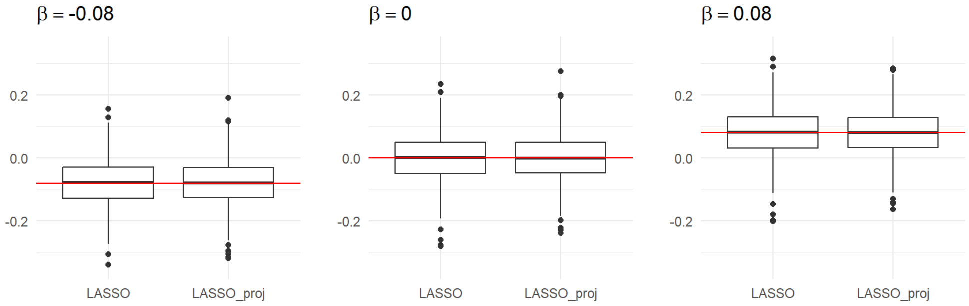


Figure A.14 Boxplots of LASSO and LASSO-projection estimators.

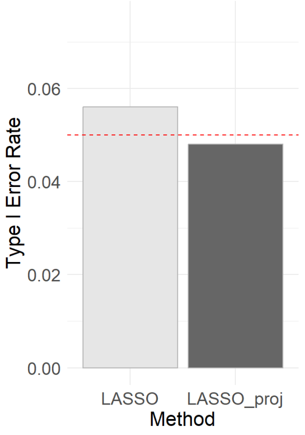


Figure A.15 Type I Error.

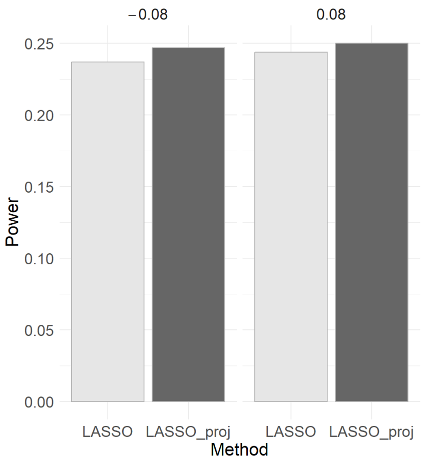


Figure A.16 Power.

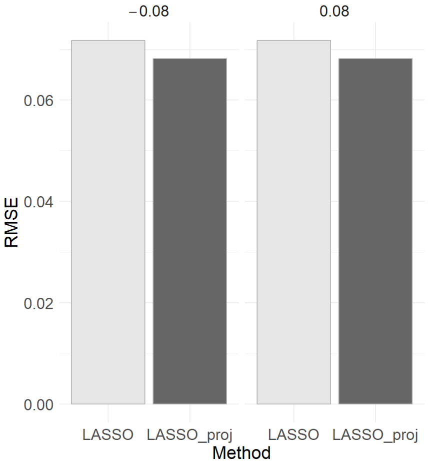


Figure A.17 RMSE.

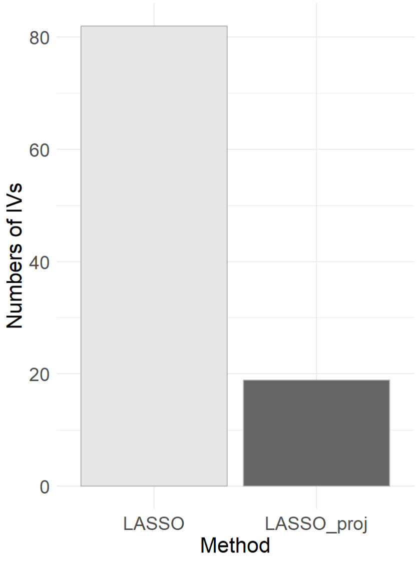


Figure A.18 Total IVs selected.

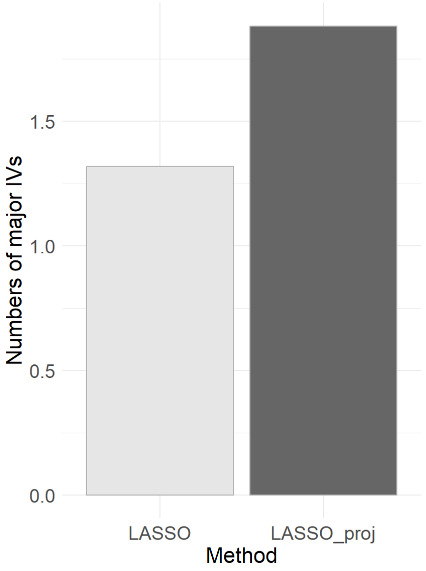


Figure A.19 Major IVs selected.

A.2.8 Additional results for the real data analysis

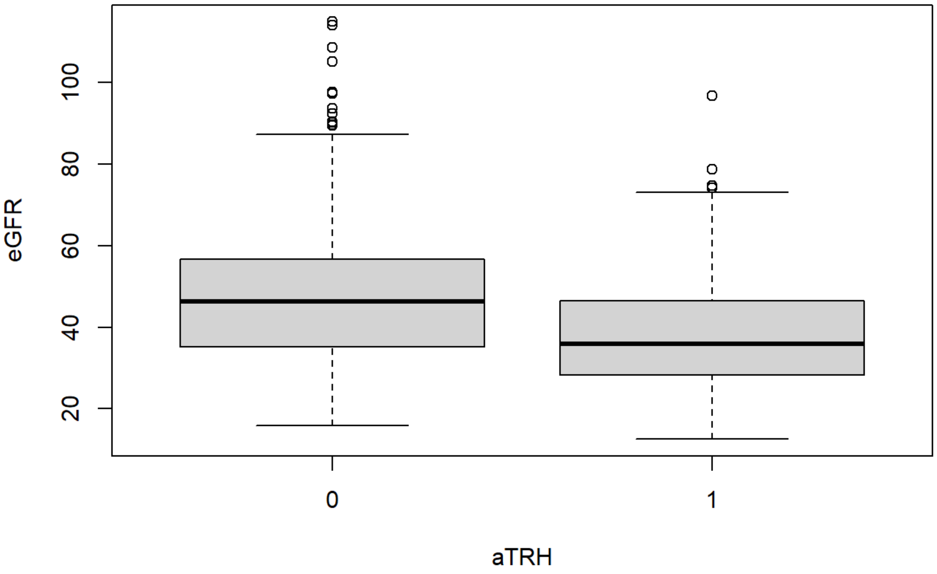


Figure A.20 Boxplot of eGFR in aTRH positive and negative groups.

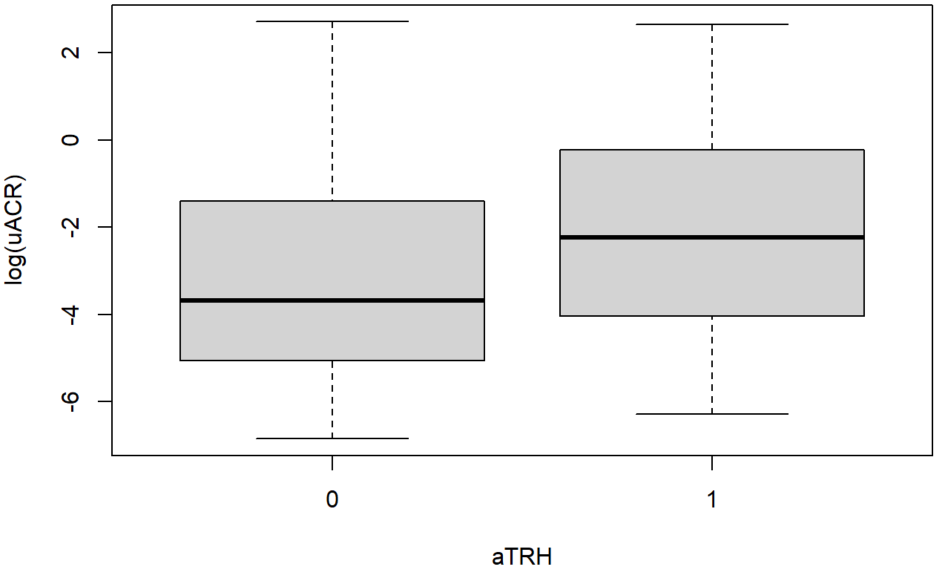


Figure A.21 Boxplot of log(uACR) in aTRH positive and negative groups.

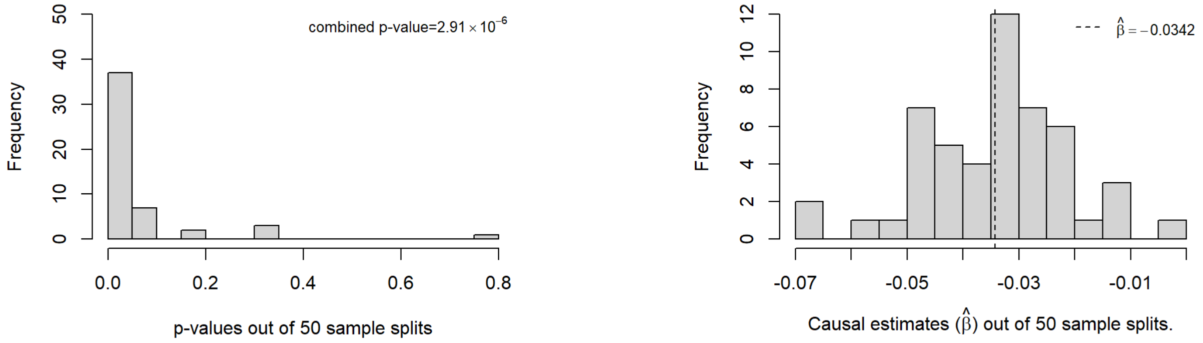


Figure A.22 Histogram of p-values and causal effect estimates from 50 sample splits when eGFR is treated as the exposure (F>20 is used to distinguish the major IVs).

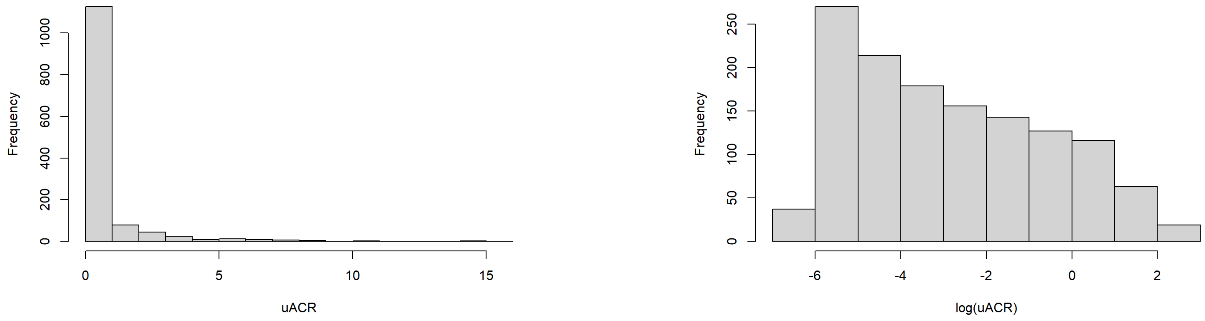


Figure A.23 Histogram of uACR (left figure) and log(uACR) (right figure).

A.3 Chapter 3

A.3.1 Supplemental figures with multi-sample splitting

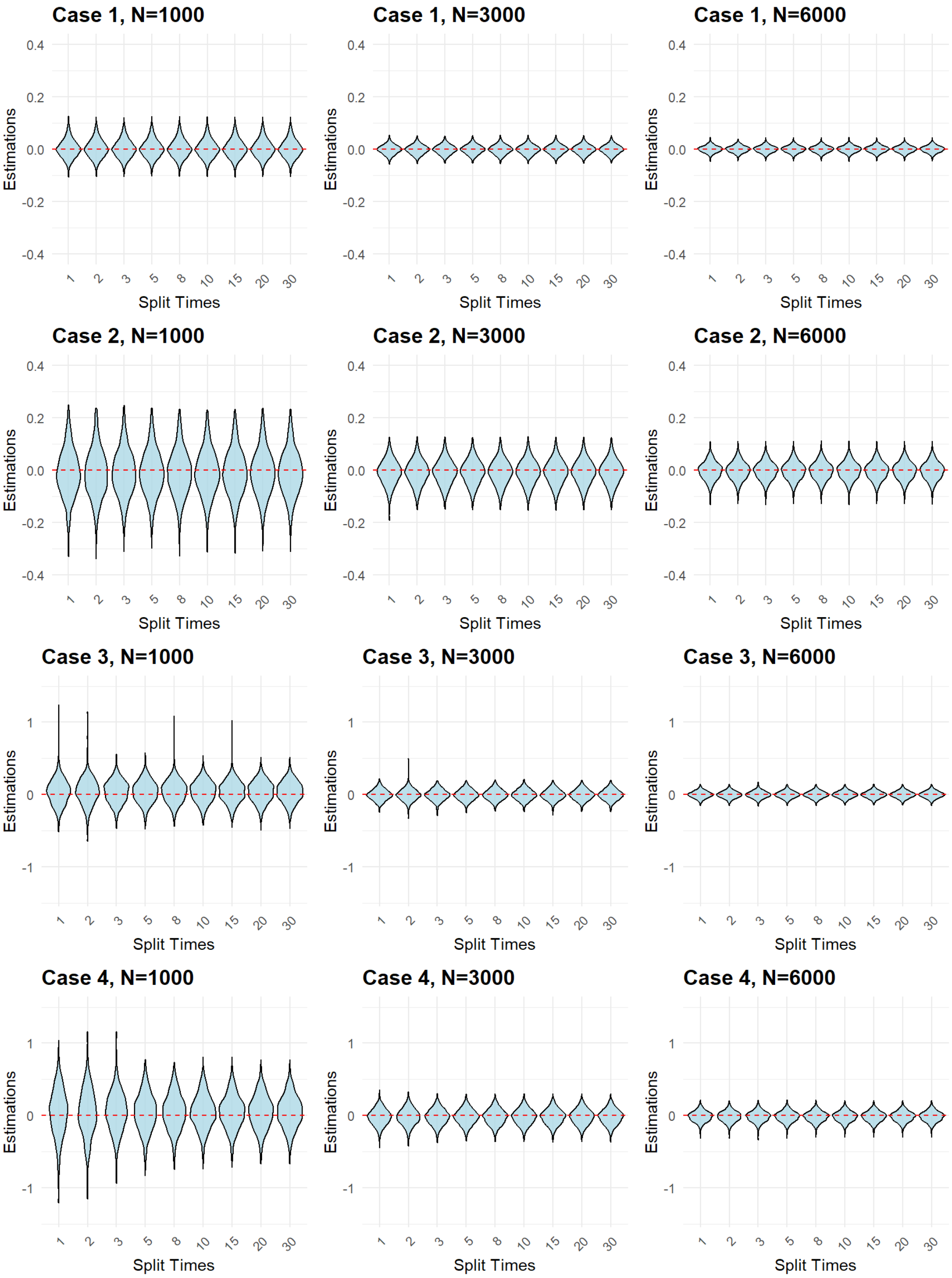


Figure A.24 Violin plots of causal effect estimates under different sample split times from 0 to 30.

A.3.2 Simulation without noise

Table A.2 The breakdown of valid IV selection results without noise.

	N	IVs	MR-SPLIT+	WIT	CIIV	sisVIVE
Case 1	1000	valid	9	7.7	9	0
		invalid	0.1	0.2	0.2	3.9
	3000	valid	9	7.9	9	0
		invalid	0	0.1	0	4
Case 2	1000	valid	9	7.7	9	0
		invalid	0.1	0.1	0.2	4
	3000	valid	9	7.9	9	0
		invalid	0	0	0	4
Case 3	1000	valid	8.8	5	8	2.4
		invalid	2.8	1.3	3.4	3
	3000	valid	9	2.9	9	0.4
		invalid	0.4	2.5	0.4	1.2
Case 4	1000	valid	8.8	4.7	7.2	1
		invalid	2.9	1.3	4.7	4.3
	3000	valid	9	7.1	9	0
		invalid	0.4	0.3	0.5	4

Note: The numbers in each row represent the average counts of being identified as valid IVs, given their true identity as valid or invalid IV, across 1,000 simulations. The true number of valid IVs is 9.

Table A.3 The breakdown of invalid IV selection results without noise.

	N	IVs	MR-SPLIT+	WIT	CIIV	sisVIVE
Case 1	1000	valid	0	1.3	0	9
		invalid	11.9	11.8	11.8	8.1
	3000	valid	0	1.1	0	9
		invalid	12	11.9	12	8
Case 2	1000	valid	0	1.3	0	9
		invalid	11.9	11.9	11.8	8
	3000	valid	0	1.1	0	9
		invalid	12	12	12	8
Case 3	1000	valid	0.2	4	1	6.6
		invalid	9.2	10.7	8.6	9
	3000	valid	0	6.1	0	8.6
		invalid	11.6	9.5	11.6	10.8
Case 4	1000	valid	0.2	4.3	1.8	8
		invalid	9.1	10.7	7.3	7.7
	3000	valid	0	1.9	0	9
		invalid	11.6	11.7	11.5	8.1

Note: The numbers in each row represent the average counts of being identified as invalid IVs, given their true identity as valid or invalid IV, across 1,000 simulations. The true number of invalid IVs is 12.

A.3.3 Simulation with noise IVs

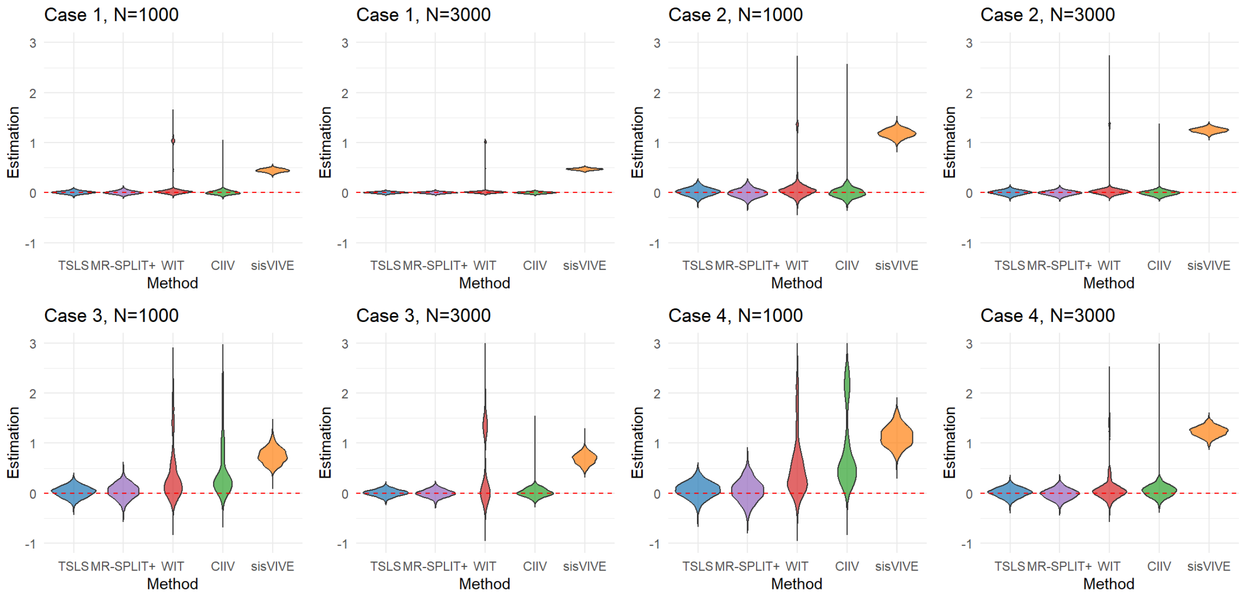
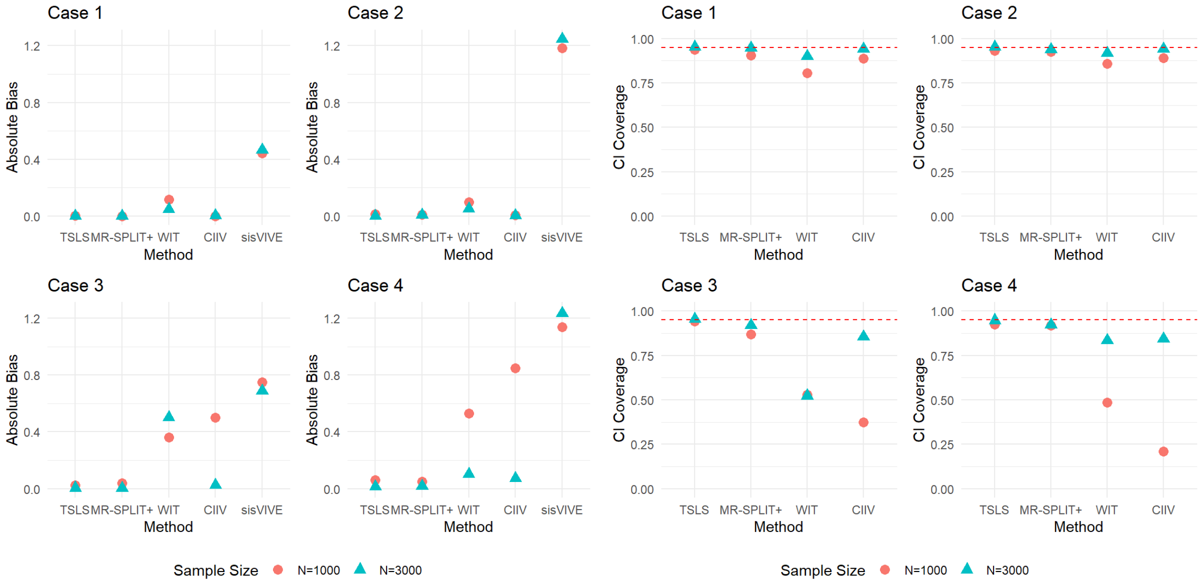


Figure A.25 Violin plots of estimators in simulations with noise.



(a) Absolute bias of estimators.

(b) Coverage Probability.

Figure A.26 Comparison of absolute bias and coverage probability for different methods in simulations with noise.

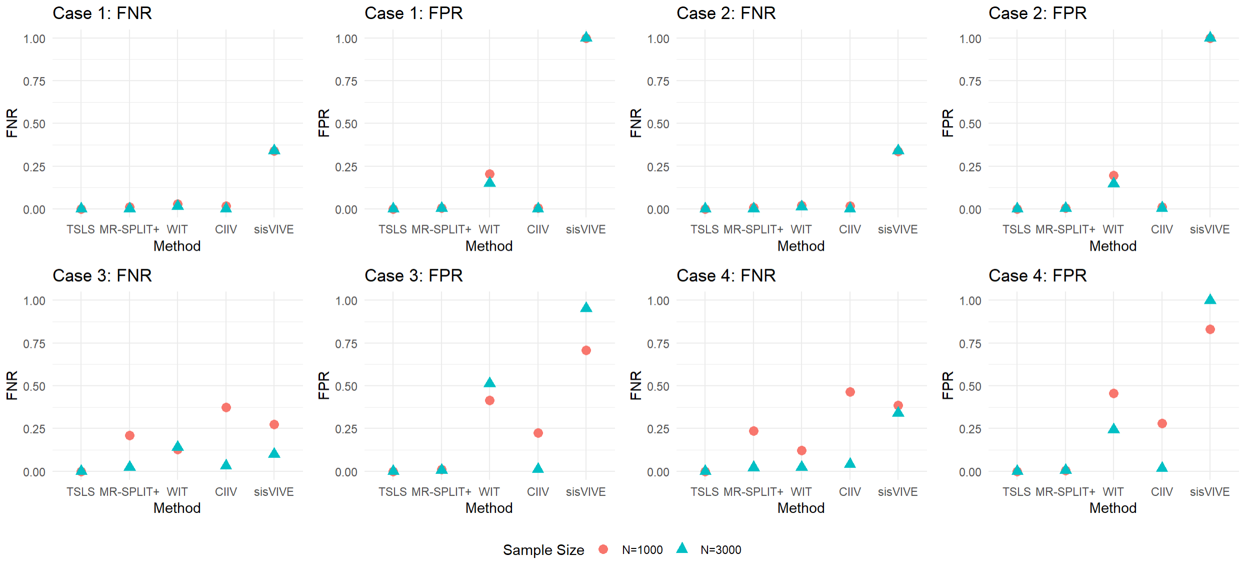


Figure A.27 Plots of False positive rate (FPR) and false negative rate (FNR) for selected IVs in simulations with noise.

Table A.4 The breakdown of valid IV selection results with noise.

	N	IVs	MR-SPLIT+	WIT	CIIV	sisVIVE
Case 1	1000	valid	9	7.2	9	0
		invalid	0.1	0.3	0.2	4.1
		noise	0.6	0.1	0.2	0.1
	3000	valid	9	7.6	9	0
		invalid	0	0.2	0	4.1
		noise	0.5	0	0.3	0.1
Case 2	1000	valid	9	7.2	8.9	0
		invalid	0.1	0.3	0.2	4
		noise	1.1	0.2	0.8	0.7
	3000	valid	9	7.7	9	0
		invalid	0	0.1	0	4.1
		noise	1.1	0.2	0.9	0.7
Case 3	1000	valid	6.8	4.6	6.2	2.3
		invalid	1.8	1.3	3.8	2.8
		noise	1.4	0.4	1.1	0.9
	3000	valid	8.9	4.4	8.9	0.4
		invalid	0.3	1.7	0.4	1.2
		noise	1.4	0.2	1	0.7
Case 4	1000	valid	5.9	4	5.4	1.3
		invalid	1.9	1.2	4.7	3.8
		noise	1.7	0.5	1.2	1.1
	3000	valid	8.9	6.8	8.8	0
		invalid	0.3	0.3	0.5	4.1
		noise	1.6	0.2	1.2	1.1

Note: The numbers in each row represent the average counts of being identified as valid IVs, given their true identity as valid, invalid, or noise IV, across 1,000 simulations. The true number of valid IVs is 9. Noise refers to variants that have no effect on either the exposure or the outcome, but are incorrectly classified as valid IVs.

Table A.5 The breakdown of invalid IV selection results with noise.

	N	IVs	MR-SPLIT+	WIT	CIIV	sisVIVE
Case 1	1000	valid	0	1.8	0	9
		invalid	11.9	11.7	11.8	7.9
		noise	0	0.2	0.1	0.2
	3000	valid	0	1.4	0	9
		invalid	12	11.8	12	7.9
		noise	0	0.3	0	0.2
Case 2	1000	valid	0	1.8	0.1	9
		invalid	11.9	11.7	11.8	8
		noise	0.1	0.7	0.1	0.2
	3000	valid	0	1.3	0	9
		invalid	12	11.9	12	7.9
		noise	0.1	0.7	0.1	0.2
Case 3	1000	valid	0.1	3.2	1.6	5.4
		invalid	6.9	8.8	6.3	7.3
		noise	0.2	0.9	0.2	0.3
	3000	valid	0.1	4.6	0.1	8.6
		invalid	11.7	10.3	11.6	10.8
		noise	0.1	1	0.3	0.5
Case 4	1000	valid	0	3.3	1.9	6
		invalid	6.2	8.6	5.2	6
		noise	0.2	0.9	0.1	0.2
	3000	valid	0.1	2.2	0.2	9
		invalid	11.7	11.7	11.5	7.9
		noise	0.1	1.1	0.2	0.3

Note: The numbers in each row represent the average counts of being identified as invalid IVs, given their true identity as valid, invalid, or noise IV, across 1,000 simulations. The true number of invalid IVs is 12. Noise refers to variants that have no effect on either the exposure or the outcome, but are incorrectly classified as invalid IVs.

A.4 Chapter 4

A.4.1 Simulation results

Table A.6 Simulation results of scenario 1 when $\beta_{XY} = \beta_{YX} = 0$.

Settings	N	Method	Bias	Est.sd	RMSE	CI Width	CP	FPR	FNR
$\beta_{XY} = 0$	1000	Oracle TSLS	0.0019	0.0219	0.0220	0.0913	0.96	0.00	0.00
		BiMR-SPLIT+	0.0495	0.0459	0.0674	0.1440	0.68	0.13	0.01
		MR-Egger	-0.2195	0.2146	0.3069	1.2850	0.98	NA	NA
		CIIV	0.9118	0.3744	0.9856	0.1960	0.11	0.63	0.88
	2000	Oracle TSLS	0.0000	0.0156	0.0156	0.0644	0.95	0.00	0.00
		BiMR-SPLIT+	0.0191	0.0370	0.0416	0.1037	0.80	0.06	0.01
		MR-Egger	-0.1965	0.1570	0.2515	1.0963	0.99	NA	NA
		CIIV	0.7027	0.7205	1.0059	0.1797	0.44	0.28	0.52
	4000	Oracle TSLS	0.0002	0.0119	0.0119	0.0453	0.95	0.00	0.00
		BiMR-SPLIT+	0.0030	0.0236	0.0238	0.0760	0.87	0.01	0.00
		MR-Egger	-0.1785	0.1361	0.2244	0.9983	1.00	NA	NA
		CIIV	0.5569	0.7739	0.9529	0.1293	0.62	0.17	0.35
$\beta_{YX} = 0$	1000	Oracle TSLS	0.0000	0.0235	0.0234	0.0915	0.95	0.00	0.00
		BiMR-SPLIT+	0.0498	0.0531	0.0728	0.1448	0.65	0.13	0.02
		MR-Egger	-0.0304	0.2090	0.2110	1.2782	0.99	NA	NA
		CIIV	0.6552	0.4053	0.7702	0.1750	0.23	0.58	0.72
	2000	Oracle TSLS	-0.0007	0.0163	0.0163	0.0644	0.96	0.00	0.00
		BiMR-SPLIT+	0.0174	0.0355	0.0395	0.1036	0.79	0.06	0.01
		MR-Egger	-0.0680	0.1778	0.1902	1.2599	1.00	NA	NA
		CIIV	0.4530	0.3492	0.5717	0.1326	0.33	0.39	0.63
	4000	Oracle TSLS	-0.0003	0.0114	0.0114	0.0453	0.95	0.00	0.00
		BiMR-SPLIT+	0.0028	0.0242	0.0244	0.0760	0.89	0.01	0.00
		MR-Egger	-0.0916	0.1688	0.1919	1.2546	1.00	NA	NA
		CIIV	0.4344	0.2843	0.5190	0.1009	0.28	0.36	0.70

Table A.7 Simulation results of scenario 2 when $\beta_{XY} = \beta_{YX} = 0$.

Settings	N	Method	Bias	Est.sd	RMSE	CI Width	CP	FPR	FNR
$\beta_{XY} = 0$	1000	Oracle TSLS	0.0019	0.0219	0.0220	0.0913	0.96	0.00	0.00
		BiMR-SPLIT+	0.0328	0.1325	0.1364	0.1539	0.89	0.10	0.03
		MR-Egger	-0.1046	0.2207	0.2440	1.3036	0.99	NA	NA
		CIIV	0.8653	0.3642	0.9387	0.1865	0.12	0.69	0.86
	2000	Oracle TSLS	0.0000	0.0156	0.0156	0.0644	0.95	0.00	0.00
		BiMR-SPLIT+	0.0304	0.1530	0.1559	0.1098	0.91	0.08	0.04
		MR-Egger	-0.0167	0.1433	0.1441	1.1288	1.00	NA	NA
		CIIV	0.8815	0.3276	0.9403	0.1345	0.11	0.60	0.89
	4000	Oracle TSLS	0.0002	0.0119	0.0119	0.0453	0.95	0.00	0.00
		BiMR-SPLIT+	0.0256	0.1338	0.1361	0.0785	0.91	0.03	0.04
		MR-Egger	0.0256	0.1114	0.1142	1.0354	1.00	NA	NA
		CIIV	0.9272	0.2663	0.9646	0.1028	0.06	0.52	0.93
$\beta_{YX} = 0$	1000	Oracle TSLS	0.0000	0.0235	0.0234	0.0915	0.95	0.00	0.00
		BiMR-SPLIT+	0.0221	0.1019	0.1041	0.1544	0.90	0.09	0.02
		MR-Egger	-0.1086	0.2330	0.2569	1.3324	0.99	NA	NA
		CIIV	0.8902	0.3334	0.9505	0.1875	0.10	0.71	0.89
	2000	Oracle TSLS	-0.0007	0.0163	0.0163	0.0644	0.96	0.00	0.00
		BiMR-SPLIT+	0.0239	0.1354	0.1373	0.1098	0.91	0.07	0.03
		MR-Egger	-0.0164	0.1412	0.1420	1.1223	1.00	NA	NA
		CIIV	0.8588	0.3567	0.9298	0.1350	0.13	0.58	0.86
	4000	Oracle TSLS	-0.0003	0.0114	0.0114	0.0453	0.95	0.00	0.00
		BiMR-SPLIT+	0.0328	0.1569	0.1602	0.0787	0.90	0.04	0.05
		MR-Egger	0.0211	0.1091	0.1110	1.0386	1.00	NA	NA
		CIIV	0.9330	0.2489	0.9656	0.1016	0.05	0.53	0.94